UNIVERSITY
of
INFORMATION TECHNOLOGY

ICAIT 2018
International Conference
on Advanced Information Technologies

Proceedings of

the 2nd International Conference on Advanced Information Technologies (ICAIT 2018)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

長岡技術科学大学
Nagaoka University of Technology

University of the Ryukyus

Saitama University
埼玉大学

MCF

Organized by

University of Information Technology
Ministry of Education, Myanmar

Yangon, Myanmar
1st & 2nd November, 2018

Proceedings of the 2nd International Conference on Advanced Information Technologies (ICAIT 2018)

Proceedings of the 2nd International Conference on Advanced Information Technologies

ICAIT 2018

1st & 2nd November, 2018

Yangon, Myanmar

Organized by

University of Information Technology
Ministry of Education, Myanmar

ISBN 978-99971-0-586-8

9 789997 105868 >

ICAIT 2018 Conference Committee Members

General Chair:

- Prof. Saw Sanda Aye, Rector, University of Information Technology, Yangon, Myanmar

Co-Chair:

- Prof. Mie Mie Thet Thwin, Rector, University of Computer Studies, Yangon, Myanmar

Organizing Committee:

- Prof. Moe Pwint, University of Computer Studies, Mandalay, Myanmar
- Prof. Win Aye, Myanmar Institute of Information Technology, Myanmar
- Prof. Thinn Thu Naing, University of Computer Studies (Taunggyi), Myanmar
- Prof. Thandar Thein, University of Computer Studies (Maubin), Myanmar
- Prof. Khin Mar Lar Tun, University of Computer Studies (Pathein), Myanmar
- Prof. Aung Win, University of Technology (Yadanarpone Cyber City), Myanmar
- Prof. Soe Soe Khaing, University of Computer Studies (Monywa), Myanmar
- Prof. Wint Thida Zaw, University of Information Technology, Yangon, Myanmar
- Prof. Myat Thida Mon, University of Information Technology, Yangon, Myanmar
- Prof. Swe Zin Hlaing, University of Information Technology, Yangon, Myanmar
- Prof. Ei Chaw Htoon, University of Information Technology, Yangon, Myanmar
- Prof. Khin Mo Mo Tun, University of Information Technology, Yangon, Myanmar
- Prof. Htar Htar Lwin, University of Information Technology, Yangon, Myanmar
- Prof. Swe Swe Oo, University of Information Technology, Yangon, Myanmar
- Prof. Myint Thuzar Tun, University of Information Technology, Yangon, Myanmar
- Prof. Khin Kyawt Kyawt Khaing, University of Information Technology, Yangon, Myanmar

Programme-Chair:

- Prof. Aung Htein Maw, University of Information Technology, Yangon, Myanmar

Programme Committee:

- Associate Prof. Atsuo Yoshitaka, Japan Advanced Institute of Science and Technology (JAIST), Japan
- Prof. Carlo Ghezzi, Politecnico di Milano, Italy
- Assistant Prof. Chaiyachet Saivichit, Chulalongkorn University, Thailand
- Associate Prof. Chaodit Aswakul, Department of Electrical Engineering, Chulalongkorn University, Thailand
- Prof. Dong Seong Kim, University of Canterbury, New Zealand
- Dr. Dong-Yong Kwak, Electronics and Telecommunications Research Institute (ETRI), Korea
- Associate Prof. Fuyuki Ishikawa, National Institute of Informatics, Tokyo, Japan
- Prof. Hiroyuki Iida, Japan Advanced Institute of Science and Technology (JAIST), Japan
- Prof. Hiroyuki Miyazaki, University of Tokyo, Japan
- Prof. Jaeyoung Ahn, Electronics and Telecommunications Research Institute (ETRI), Korea
- Dr. Jeong Dae Suh, Electronics and Telecommunications Research Institute (ETRI), Korea
- Prof. Jong Sou Park, Korea Aerospace University, Korea
- Associate Prof. Kiyoaki Shirai, Japan Advanced Institute of Science and Technology (JAIST), Japan
- Prof. Koichiro Ochimizu, University of Information Technology, Yangon, Myanmar
- Associate Prof. Takashi Komuro, Information and Computer Sciences, Saitama University, Japan
- Associate Prof. Ling Teck Chaw, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
- Associate Prof. Lunchakorn Wuttisittikulkij, Chulalongkorn University, Thailand

- Prof. Masashi Unoki, Japan Advanced Institute of Science and Technology (JAIST), Japan
- Prof. Nobuo Funabiki, Dept. of Electrical and Communication Engineering, Okayama University, Japan
- Assistant Prof. Olarik Surinta, Department of Information Technology, Mahasarakham University, Thailand
- Prof. Tetsuya Shimamura, Information and Computer Sciences, Saitama University, Japan
- Prof. Yoshinori Sagisaka, Waseda University, Japan
- Associate Prof. Yuichi Otsuka, Nagaoka University of Technology, Japan
- Prof. Yutaka Ohsawa, Information and Computer Sciences, Saitama University, Japan
- Associate Prof. Yuto Lim, Japan Advanced Institute of Science and Technology (JAIST), Japan
- Dr. Yan Lin Aung, Singapore University of Technology and Design, Singapore
- Dr. Ye Kyaw Thu, Researcher, Okayama Prefectural University (OPU), Japan
- Dr. Yin Minn Pa Pa, Researcher, Cyber Security Laboratory, PwC Cyber Services LLC, PwC Japan
- Prof. Kalayar Myo San, University of Computer Studies, Mandalay, Myanmar
- Prof. Khaing Moe Nwe, University of Computer Studies, Yangon, Myanmar
- Prof. Khin Mar Soe, University of Computer Studies, Yangon, Myanmar
- Prof. Khin Nweni Tun, University of Computer Studies (Taunggyi), Myanmar
- Prof. Khin Than Mya, University of Computer Studies, Yangon, Myanmar
- Prof. Khin Thida Lynn, University of Computer Studies, Mandalay, Myanmar
- Prof. May Aye Khine, University of Computer Studies, Yangon, Myanmar
- Prof. Mie Mie Su Thwin, University of Computer Studies, Yangon, Myanmar
- Prof. Nang Saing Moon Kham, University of Computer Studies, Yangon, Myanmar
- Prof. Nyein Aye, University of Computer Studies (Hpa-an), Myanmar
- Prof. Nyein Nyein Myo, University of Computer Studies, Mandalay, Myanmar
- Prof. Sabai Phyu, University of Computer Studies, Yangon, Myanmar

Proceedings of the 2nd International Conference on Advanced Information Technologies (ICAIT 2018)

November, 2018

Contents

Keynote Speech

Artificial Intelligence and Robotics

Cloud and Distributed Computing

Data Mining

Database and Big Data Analytics

Image Processing

Natural Language Processing

Simulation and Modeling in Software Approach

Wireless and Software Defined Networking

Keynote Speech

Keynote Speech

Biography

Carlo Ghezzi is an ACM Fellow (1999), an IEEE Fellow (2005), a member of the European Academy of Sciences and of the Italian Academy of Sciences. He received the ACM SIGSOFT Outstanding Research Award (2015, the Distinguished Service Award (2006), and the 2018 TCSE Distinguished Education Award from IEEE Computer Society Technical Council on Software Engineering (TCSE). He has been President of Informatics Europe. He has been a member of the program committee of flagship conferences in the software engineering field, such as the ICSE and ESEC/FSE, for which he also served as Program and General Chair.

He has done research in programming languages and software engineering for over 40 years and has been a recipient of an ERC Advanced Grant on self-adaptive software systems. He has published over 200 papers in international journals and conferences and co-authored 6 books.

Software engineering for Cyber-Physical Spaces

Advances in technology, in particular in cyber-physical systems, increasingly enable functionalities that will lead to the development of smart living spaces--from "smart homes" to "smarty cities"--where living conditions for people will be facilitated and enhanced. These systems may be collectively called cyber-physical spaces. They will assist and cooperate with people in their homes, including elder and disabled people. They will enhance operations in public buildings, such as hospitals or courts. They will make public spaces more secure; e.g., airports or stations. They will assist in managing traffic in cities, reducing air pollution, and reducing energy consumption. Needless to say, the design of such cyber-physical spaces is a multidisciplinary endeavour, ranging contributions from Internet-of Things to software engineering to civil engineering and architecture to medical sciences, transportation science, environmental science, energy.

The talk will argue that software engineering can bring a unique and fundamental contribution into this multidisciplinary world. It can support the design phase with formal models that integrate existing spatial design notations (such as BIM used by architects and civil engineers; or CityGML, an emerging notation for city and landscape models) with modelling notations that support automatic reasoning and analysis, for example to check compliance with existing regulations or possible security and safety threats, or simulate how the space being designed will behave when operational. Automatic checking of models can also support monitoring the smart space, when it will be operational, and possible automatic reactions to keep the operational smart space aligned with its requirements.

Initial research results in this direction will be presented, along with a possible research agenda. The talk is a call for an interdisciplinary approach to address the problem and presents a discussion of the crucial role that software engineering can have in this area.

Keynote Speech

Biography

Professor Dr. Hiroyuki Iida is a Japanese computer scientist and computer games researcher with focus on Game-refinement theory, Opponent Model Search and Computer Shogi. Hiroyuki Iida is Professor at Japan Advanced Institute of Science and Technology (JAIST) and director of entertainment science center and entertainment technology. Before, he was affiliated with the Shizuoka University, Hamamatsu. He received his Ph.D on Heuristic Theories on Game-Tree Search from Tokyo University of Agriculture and Technology, Tokyo.

Hiroyuki Iida is a professional 7-dan Shogi GM player, and co-author of the Shogi program TACOS, the four-time Gold medal winner at Computer Olympiads. Dr. Hiroyuki Iida has been an enthusiasm researcher in the domains such as computer games and entertainment computing, while acting as important roles of international activities such as conference chair, program chair and journal editor. He has also organized Mind Sports Computer Olympiad as the secretary/treasurer of ICGA (International Computer Games Association) for each year since early 2000. He supervised many master/PhD students until now, while acting as PhD committee member (external assessment) for PhD candidates in western countries such as Maastricht University and Tilburg University in the Netherlands. He also served as an external assessment for international research funding in western countries such as North America and the Netherlands.

Deep Blue, Tacos, AlphaGo, and More

In this talk, a historical overview of computer games development is given from the viewpoint of *Using games as testbed for AI research*. Then new challenges in this direction are introduced. One is a research question "*Can a computer enjoy game playing?"*, whereas another one is "*Is it possible to visualize the thinking process of strong game AI?*" Moreover, our recent work called force-in-mind theory is introduced. It is expected that the theory enables us to better understand in-depth concept of emotion in game playing and power of gaming elements in non-game contexts.

Keynote Speech

Cyber Threat Trend and Cyber Exercises

Professor Yoichi Shinoda
School of Information Science, Security and Networks Area
Japan Advanced Institute of Science and Technology

Degrees:
B.E., M.E. and Ph.D. from Tokyo Institute of Technology (1983, 1985, 1989)

Professional Career:
Associate at Tokyo Institute of Technology (1988)
Professor of School of Information Science at JAIST (1991)

Specialties:
Distributed and Parallel Computing
Networking Systems
Operating Systems
Information Environment

Artificial Intelligence and Robotics

Advantage of Initiative Revisited: A case study using Scrabble AI

Htun Pa Pa Aung
Entertainment Technology
School of Information Science
Japan Advanced Institute of Science and Technology
Email:htun.pp.aung@jaist.ac.jp

Hiroyuki Iida
Entertainment Technology
School of Information Science
Japan Advanced Institute of Science and Technology
Email:iida@jaist.ac.jp

***Abstract*—This paper explores the advantage of initiative using Scrabble as a test bed. Recently, a list of solved two-person zero-sum games with perfect information has increased. Among them, most of the games are a win for the first player (i.e., the advantage of initiative), some are draws, and only a few games are a win for the second player. Self-play experiments using Scrabble AIs were performed in this study. The results show that the player who established an advantage in the early opening took higher win expectancy. This implies that the advantage of initiative should be reconsidered to apply for all levels including nearly perfect players. Thus, we meet a new challenge to improve the rules of a game to maintain the fairness. The game of Scrabble gives an interesting example while giving a randomized initial position. This discussion can be extended to other domains when AI becomes much stronger or smarter than before.**

***Index Terms*—Artificial Intelligence, Advantage of Initiative, Fairness, Scrabble**

I. Introduction

Game has been as an ideal test bed for the study of Artificial Intelligence (AI). Until recently, most of the academic works in the area focused on traditional board games and card games, the challenge is to beat expert human players [10]. AI can be applied to most aspects of game development and design including automated content creation, procedural animation, adaptive lighting and intelligent camera control. While most games are developed to be fun to play, there exists a class of games developed as simplified models for the study of economics or social behavior and even for the improvement of the quality of education.

Now AI can utterly dominate humans within the world of gaming. One proof was demonstrated in chess with Deep Blue [5] beating the world champion Garry Kasparov in 1997. AlphaGo [13] showed its significant performance being much better than humans. Moreover, poker AI LIBRATUS convincingly beat four professional players [1]. In this study we have chosen Scrabble as the main test bed. Scrabble [2] is a board scoring game in which players place words with tiles with different scores using own rack onto a board. It is originally played by two or more players on a 15x15 grid of cells. It has been played for decades in various situations, for instance, as a competitive match between professional players or a friendly match among family members or students. Different players have different vocabulary knowledge and supposed to play to get different experiences.

This paper concerns about "advantage of initiative" in games, which was discussed by Uiterwijk and Van den Herik [16] with respect to Singmaster's reasoning [14]. The state of the current knowledge is that many games are a win for the first player, some games are draws, and only a few games are a win for the second player. The results indicate that having the initiative is a clear advantage under the condition that the board size is sufficiently large [17]. Van den Herik et al. [18] observed that in relatively many games on small boards the second player is able to draw or even to win. However, these observations were made under the assumption that the performance level is perfect or human-like (beginners to grandmasters). Thus, our research focuses on the level of more than human but not perfect. A question may arise: "What will the game-theoretic value be if both players are stronger than human but not perfect?" Since AI in games becomes (much) stronger than human, it is possible to find the answer.

The present contribution is expected to enhance the completeness of initiative from different perspective. Then, we focus on the advantage of initiative from the perspective of very strong players using Scrabble as a test bed. We conduct self-play experiments using Scrabble AIs of different levels to observe the impact of the initiative. Fairness or equality is another essential aspect of games. Without it, a game would lose its charm and therefore be forgotten in the past. So, it is a serious matter to maintain fairness as well as attractiveness in the history of the games [6]. Another question may arise: "Which factor should be considered to maintain fairness for those who are stronger than human but not perfect?".

The course of the paper is as follows. Section II describes the basic overview of Scrabble and early work of Scrabble. Section III discusses the advantage of initiative from AI's perspective. Experimental evidence on playing Scrabble by AI is described in Section IV. Finally, Section V gives conclusions and future work.

II. Scrabble and its Early Works

A. *SCRABBLE*

Scrabble is a game with a long history, which still maintains its popularity without major changes in its regulations. Scrabble is an imperfect information game but becomes a perfect information game during the endgame phase. Scrabble is affected by the chance factor during the draw phase. In general, Scrabble is luck-dependent during the draw phase. Thus, both players cannot predict for a future draw. So, it is more effective to play with a local best move. Each move may lead to a different outcome. However, Scrabble is such a kind of two player scoring game played on the board.

B. Early works with Scrabble

Scrabble has been an interesting target for search algorithm to develop the fastest one. Among many works, for example, see [4]. Another direction is to explore the entertaining aspect of playing Scrabble. Scrabble was analyzed using game refinement measurement. For example, the sophistication of Scrabble was assessed from the viewpoint of entertainment [8]. The results indicate that the game refinement value of Scrabble is slightly higher than the zone of well sophisticated games such as chess. We show, in Table I, the measures of game refinement for several games.

TABLE I
MEASURES OF GAME REFINEMENT FOR VARIOUS GAMES

Game	**GR**
Scrabble [8]	0.083
Chess [7]	0.074
Go [7]	0.076
Basketball [15]	0.073
Soccer [15]	0.073

The swing model, a derivation of the game progress model, is defined to solve the nonidentical scoring system in Scrabble [8]. Swing denotes a notion of phase transition in mind from advantage to disadvantage and vice versa. In previous work [8], a computer program was built to simulate multiple Scrabble matches to estimate the game refinement value.

III. ADVANTAGE OF INITIATIVE AND GAME-THEORETIC VALUE

A two player game with a turn to move would not be enjoyable if it cannot keep fairness of the winning ratio [6]. It is observed in the previous work [7] that it is more interesting for players to play a game in which the information about the game outcome is not clear at the very end of the game than to play a game where the outcome is already determined after a few rounds or stages. An important factor is that all players must feel fair in the game so that the game can maintain its attractiveness besides keeping some degree of competitiveness.

A. Advantage of the initiative

During the last decade, several two-person zero-sum games with perfect information have been solved [16] [18]. The state of current knowledge is that many games are a win for the first player, some games are draws, and only a few games are a win for the second player. Uiterwijk and Van den Herik [16] distinguished two main concepts valid for any two players games, namely initiative and zugzwang. The initiative was defined as an action of the first player. The results from their experiments show that having the initiative is a clear advantage under the condition that the board size is sufficiently large. With respect to Singmaster's reasoning [14] Van den Herik et al. [18] observed that in relatively many games on small boards the second player is able to draw or even to win. Thus, it can be assumed that Singmaster's reasoning has limited value when the board size is small.

On the other hands, Kita and Iida [9] studied a link between the initiative and the game-theoretic value with a focus on the mobility in the initial position. The results of the exhaustive analysis of possible initial positions and game-theoretic values in the domain of 4x4 reversi show that the game-theoretic value is positively correlated with the mobility in the initial position.

Of all games solved, many are the first-player win [16]. These games show that having the initiative is an advantage. Therefore, it is worth investigating what happens if the initiative fails. In Table II, we have collected the main results known today [16] [18], in which 0 stands for the draw, 1 for a first player win, and 2 for a second player win.

TABLE II
SOME GAME-THEORETIC VALUES OF GAMES KNOWN TODAY

Game	**Result**
Connect-Four	1
Qubic	1
Nine Men's Morris	0
1xm Go (m =1,2,5)	0
1x3, 1x4, 2x4, 3x3, 3x4 Go	1
6x6 Othello	2
mnk-games(k=1,2)	1
333-game(TicTacToe)	0
mn3-games($m \geq 4, n \geq 3$)	1
m44-games($m \leq 8$)	0
mn4-games($m \leq 5, n \leq 5$)	0
mn4-games($m \geq 6, n \geq 5$)	1
mn4-games($m \leq 5, n \leq 5$)	0
mn5-games($m \leq 6, n \leq 6$)	0
19, 19,5-game (Go Moku)	1
mnk-games($k \geq 8$)	0
mxm Domineering (m=1,5)	2
mxm Domineering (m=2,3,4,6,7,8)	1

B. Initiative in Scrabble

Scrabble is a scoring game played on the board which is slightly different from standard board games. According to its basic regulations and history, there is no fixed initial position in Scrabble. Letters are randomly distributed to each player in the initial stage of the game. In this paper, we investigate the impact of the advantage of early stages in Scrabble under the condition that the level of players are more than human but not perfect. In Scrabble, the initiative is defined as the action of each player in the first stage. To investigate the impact of the initiative in Scrabble, we set up experiments on Scrabble AI to collect data as well as from Scrabble human tournaments.

IV. Experiments and Results

A. Human expert Scrabble players

1) Human Scrabble strategy: Human Scrabble strategy can roughly be defined according to the following four phases:

1) search for a bingo
2) search for hot spots
3) try to improve upon the results found, and
4) consider the rack leave

First, the player should always try to find a bingo by looking for either a 7-letter word using the entire rack or an 8-letter word that uses one tile from the board. Second, he can look for hot spots including premium squares which are different types, e.g., doubling and tripling the score of a letter and also doubling and tripling the score of a word. Third, he tries to get much advantage upon the previous result of the board. Finally, the player should consider the remaining rack leave for future steps because sometimes the current highest scoring move is not the best option and it can leave bad tiles in the rack for next move.

2) Tournaments results of human expert players: Normally, human players use a very common strategy to open the board with high scores. Such players would always want to keep the board open since they are more able to utilize the board's openness than their opponent. On the other hand, most human players use several strategies simultaneously. According to human tournament results, even the human experts have higher chance to be winner of the game when he took higher score as much as possible than his opponent in the first turn of the game.

TABLE III
Early stage advantage and disadvantage compared in human expert tournaments (N=6000)

	Advantage	Disadvantage
Win ratio	4200 (70%)	1800 (30%)

B. Scrabble AI – Quackle

Almost all of the improvements were driven by breakthroughs in artificial intelligence growing ability to understand complex nuances of the world around it. Quackle is not only a Scrabble Game but also an artificial intelligence and analysis tool. There are three key features in this AI from the perspective of artificial intelligence and analysis [11]. First, all the playable moves from the game current position. Second, Quackle can run a simulation, by playing itself hundreds of times, and tell the player how often each of these moves ends up winning. Third, the player can find out how many other mistakes the player made during the game.

1) How Quackle *works:* Computer Scrabble Quackle was used to simulate multiple scrabble matches with two AI players. The results with essential data with individual scores and total scores are collected. The database of estimated win probabilities was implemented by analyzing the distribution of wins over many Quackle self-play games. This kind of win-percentage-based analysis is critical in a Scrabble AI when we need to erase a large deficit or protect a lead.

2) Experiment results on Quackle: In this section, the results of the experiment performed with Quackle are discussed. Although Player A and Player B have the same strength and they are almost perfect players, the player who got the advantage in the initial stage of the game has the higher probability to be a winner of the game. As shown in Figure **??**, there are two Quackle AI with the same strength and Player B got an advantage at the early stage, so Player A has very low probability to win the game. At the final stage, Player A won the game because of the advantage in the first stage. We made Scrabble match simulation with Quackle. The players are almost perfect. The results are shown in Figure 1. No matter who got played first in Scrabble, the result in the first stage can guess the probability of winning fraction of scrabble. In Figure 1, there are total 60 matches by Quackle and from the results, the probability of winning fraction of the match can be estimated as shown in Table IV.

TABLE IV
Early stage advantage and disadvantage compared in 6000 games played by Quackle

	Advantage	Disadvantage
Win ratio	4400 (73.3%)	1600 (26.7%)

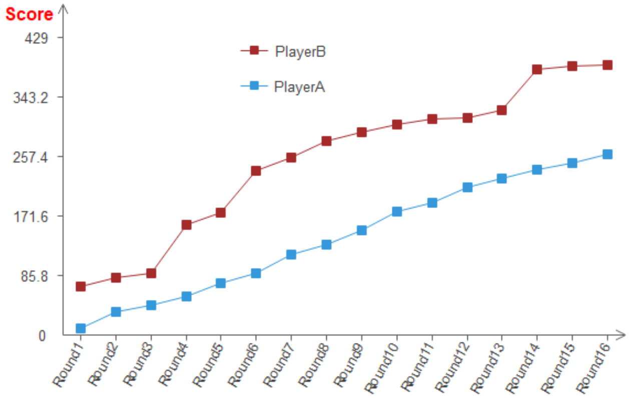


Fig. 1. A self-play experiment using Quackle: score at each turn

According to the experiment, only 11 out of 2566 games were tied (0.4 %), player 1 won 1404 games compared to Player 2s 1151 games. We can test if this is a significant difference using a binomial test.

C. Scrabble AI – Maven

Maven [12] is another Scrabble AI, created by Brian Sheppard. It has been used in official licensed Hasbro Scrabble games. In 1983, Maven's first version was better than anything, but it was nothing special. By 1998, Maven had very good methods of controlling CPU utilization so that it could play at interactive speeds. This was the first commercial product that a

Perfect level. And then Maven Champion level was developed, it is probably stronger than human champions by seeing the statistics of tournaments' results.

1) How Maven plays: Maven is divided into three sub-phases: mid-game phase, pre-endgame, and endgame. Maven performs three basic operations to make the best move: move generation, rack evaluation and search and evaluation and simulation. Maven uses different search engines, each specializing in one phase of the game. One of the challenges of programming Maven was to produce the realistic play at the weak levels. Maven has a common-word dictionary that it employs at levels below Champion, and move generation strategies that result in low average scores without resorting to frequent exchanges.

As the primary goal of heuristic is to find the best matching solution of the original answer, Maven uses four main heuristics to select the most promising moves for a player [3]. The first one is called Vowel-Consonant for balanced rack management to examine if the rack has a right mix of Vowels and Consonants. The second is known as U-with-Q-Unseen to give a priority to play the words that contain a combination of Q and U. The third heuristic is called Hot-Spot Block where the board-square near the premium squares are blocked by the player in the current turn. The fourth one is First-Turn-Open that implies the importance of playing the first turn with a few tiles.

2) Maven tournament statistics: Maven plays better than the human experts [12], as shown in Table V. According to [12], Maven's total matches and tournament record is 3500 wins and 1500 losses against an average rating of 1975. Finally, what qualities of MAVEN make it such a very strong player? The reason is that MAVEN knows all the words, it evaluates the moves well, it plays fast that makes the other player into time trouble, it never loses a challenge and it plays endgames perfectly. MAVEN is a machine, so it never gets tired and inattentive. That is one of the significant factors in game play. Because of these skills, it becomes a formidable opponent for human players.

The biggest challenge in the world of Scrabble AI is met because no one disputes that MAVEN is better than any human seeing the statistics in Figure V.

TABLE V
MAVEN AND HUMAN EXPERTS COMPARED [12]

	MAVEN	Human expert
Average Bingo per game	1.9	1.5
Average tiles played per game	4.762	4.348
Average turns per game	10.5	11.5
Chance to play Bingo if exists	100%	85%

TABLE VI
EARLY STAGE ADVANTAGE AND DISADVANTAGE COMPARED IN MAVEN SCRABBLE TOURNAMENTS (N=5000)

	Advantage	Disadvantage
Win ratio	3500 (70%)	1500 (30%)

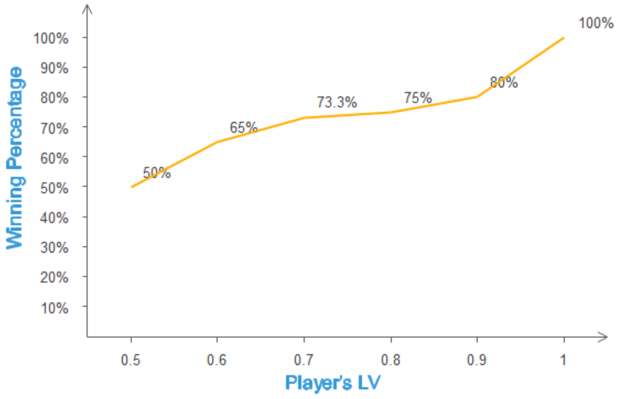


Fig. 2. Impact of Player's Strength on Winning Percentage

V. Conclusion and Future Work

Scrabble is the main test-bed of this study. We used our AI intelligent player, MAVEN, QUACKLE and human tournament statistics to deeply comprehend its special characteristics. The notion of initiative introduced by Singmaster [14] can be considered for players of normal level from beginners to grandmasters. While AI is becoming stronger than human and even at nearly perfect level, we revised the earlier model from the perspective of different player's level (from intermediate to perfect) in Scrabble game. Singmaster (1981; 1982) showed a reasoning of why first-player wins should abound over second-player wins. However, Van den Herik et al. (2002) observed through the exhaustive computer analysis that in relatively many games on small boards the second player is able to draw or even to win [18]. Hence, it is assumed that the Singmasters reasoning has limited value when the board size is small.

From an investigation of solved games the concept of initiative seems to be a predominant. Thus, in our study, we observed that if the player got advantage in the first stage of the game, the initial position of a given game would take an advantage of the initiative and he has also higher winning percentage. The higher the player's level is, the greater winning percentage he takes. However, it is supposed that although real match between novice players is difficult to obtain if the level of players are not very strong and they have different strength, we could say that they probably have equal chance to win the game. Moreover, the decision complexity of both of human players is almost the same. But for AI players, the decision complexity of winner is slightly higher than that of a loser.

There is still an important issue, i.e., all participants must feel fair in the game. The next step is to consider fairness issue for AI players in Scrabble game. Moreover, one possible way is to enhance the original rules of Scrabble which would make both AI players to get fairness. Clearly, it will be more interesting for players to play a game in which the information about the game outcome is not clear until the very end of the game. For future work, fairness needs to be considered not only in Scrabble but also in other domains as well. It is an important factor not only for games but also in society. Further verification and investigation are also left for future work.

Acknowledgment

This research is funded by a grant from the Japan Society for the Promotion of Science, within the framework of the Grant-in-Aid for Challenging Exploratory Research.

References

[1] Ai dominating humans in the world of gaming. https://futurism.com/five-examples-of-ai-dominating-humans-in-the-world-of-gaming/. (Visited on 06/06/2017).

[2] Scrabble history :making of the classic american board game. http://scrabble.hasbro.com/en-us/en-us/history. (Visited on 06/05/2017).

[3] Priyatha Joji Abraham. A scrabble artificial intelligence game. Master's thesis, San Jos State University, 2017.

[4] Andrew W. Appel and Guy J. Jacobson. The world's fastest scrabble program. *Commun. ACM*, 31(5):572–578, May 1988.

[5] Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1):57 – 83, 2002.

[6] H. Iida. On games and fairness. In *12th Game Programming Workshop in Japan*, pages 17–22, 2007.

[7] Hiroyuki Iida, Nobuo Takeshita, and Jin Yoshimura. A metric for entertainment of boardgames: its implication for evolution of chess variants. In *Entertainment Computing*, pages 65–72. Springer, 2003.

[8] S. Kananat, J.-C. Terrillon, and H. Iida. Gamification and scrabble. In Rosa Bottino, Johan Jeuring, and Remco C. Veltkamp, editors, *Games and Learning Alliance*, pages 405–414, Cham, 2016. Springer International Publishing.

[9] Hayato Kita, Cincotti Alessandro, and Hiroyuki Iida. Theoretical value prediction in game-playing. *SIG-GI, IPSJ (GI)*, 2005(87):71–77, sep 2005.

[10] S. M. Lucas. Computational intelligence and ai in games: A new ieee transactions. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(1):1–3, March 2009.

[11] Mark Richards and Eyal Amir. Opponent modeling in scrabble. In *IJCAI*, 2007.

[12] B. Sheppard. World-championship-caliber scrabble. *Artificial Intelligence*, 134(1):241 – 275, 2002.

[13] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

[14] D. Singmaster. Almost all games are first person games. *Eureka*, 41:33–37, 1981.

[15] Arie Pratama Sutiono, Rido Ramadan, Peetikorn Jarukasetporn, Junki Takeuchi, Ayu Purwarianti, and Hiroyuki Iida. A mathematical model of game refinement and its applications to sports games. *EAI Endorsed Transactions on Creative Technologies*, 15(5), 10 2015.

[16] J. W. H. M. Uiterwijk and H J. van den Herik. The advantage of the initiative. *Information Sciences*, 122(1):43–58, 2000.

[17] J. W. H. M. Uiterwijk and H. J. van den Herik. The advantage of the initiative. *Inf. Sci.*, 122(1):43–58, January 2000.

[18] H.Jaap van den Herik, Jos W.H.M. Uiterwijk, and Jack van Rijswijck. Games solved: Now and in the future. *Artificial Intelligence*, 134(1):277 – 311, 2002.

Development of a Software Program for Automatic Cartesian Farming Robot

[1]Jedsada Tangmongkhonsuk, [2]Abhishek Meena, [1,3]Pruk Sasithong, [1]Sanika Wijayasakra,
[1,3]Gridsada Phanomchoeng, [1,3]Chairat Phongphanphanee, Pisit Vanichchanunt[4] and
[1,3]Lunchakorn Wuttisittikulkij
[1]Faculty of Engineering, Chulalongkorn University
[2]Faculty of Engineering, Indian Institute of Technology Ropar
[3]Smart Wireless Communication Ecosystem Research Group, Chulalongkorn University
King Mongkut's University of Technology North Bangkok
Email: gridsada.phanomchoeng@gmail.com, jedsada.tang@gmail.com

Abstract

This paper presents the development of a software platform for our automatic Cartesian farming robot hardware prototype, which is capable of growing multiple types of crops in the soil bay. Seeding and plant watering are carried out automatically and efficiently throughout the crops life cycle virtually without using human labor. Users can design their own plan of growing crops in the farming space through implemented software application from anywhere, anytime using a computer notebook or desktop. Commands from users are then sent among sever which is the Raspberry Pi board and client desktop or laptop device using G-code. This prototype is particularly useful for a home or a small organization to grow crop with more controlled environment and careful management.

Key Words: Automation, Farming Robot, G-code

1. Introduction

Urbanization refers to the population shift from rural to urban residency, the gradual increase in the proportion of people living in urban areas, and the ways in which each society adapts to this change. It is predicted that by 2050 about 64% of the developing world and 86% of the developed world will be urbanized [1].

The rapid economic growth has influenced the larger number of populations to shift from rural to urban residency which is referred to as urbanization. In 2030, the number of urban areas will increase about 60% [1] and this change will create a numerous problem to life style of human being such as urban poverty, urban food insecurity, changes in the climate, crisis in natural resources and many more.

Urban agriculture gives an excellent strategy to reduce urban poverty and food insecurity and further develop the urban environmental management. Urban agriculture plays an important role in enhancing urban food security since the costs of supplying and distributing food to urban areas based on rural production continuous to increase and do not satisfy the demand, especially of the poorer sectors of the population. Therefore, it is especially required use the technology to have a small-scale modular which can be use in an indoor environment to control the growing environment.

FarmBot [2], a commercially available robotic open hardware system, offers a new way to grow multiple types of crops in a small area. Seeds are precisely planted in the soil and crops are automatically watered depending on their needs while the weeds are constantly eliminated. To control and configure FarmBot, there is a free FarmBot web application at my.farm.bot. Thousands of people around the globe have grown crops in their backyards using FarmBot systems.

In our previous work [3], we have built our own automatic Cartesian farming robot hardware prototype that works with the cloud-based software platform of FarmBot. Although the FarmBot can be controlled from web browser using a PC, in some scenarios it is not possible to provide a continue internet connection to the FarmBot. Then it doesn't receive any commands from the FarmBot web application and would not be able to follow the given commands by the user [4]. The main aim of this development project is to build software platform using python from the scratch which is compatible with our FarmBot hardware platform. By using this platform, it will be convenient to control the FarmBot for developers who want to use it in an offline.

The rest of the paper is structured as follows. Section 2 describes overall system and Section 3 explain about implemented software in detail. In the Section 4 we discussed the obtained results and we conclude our work in Section 5.

2. Overall system

In this section the overall hardware prototype of FarmBot system is given in detail. This prototype can be divided into three main components namely, mechanical, electrical and control system. As shown in Figure 1, the FarmBot prototype uses in this work is in 1 meter in length, 0.7 meter in width and about 0.8 meter in overall height. The mechanical system includes three movable axes consist of X, Y and Z axes. X and Y axes use a mechanical system called "Pinion Belt" which can use to fix the belt with the track and allow motors move with the construction along the track. For Z axis the "Lead Screw" mechanism is used which allows tools to move precisely. At the lower end of the Z axis, the universal tool mount is fixed. Our tool set consists of a seeder tool and a watering tool. It can be moved around the farming area according to the control order received.

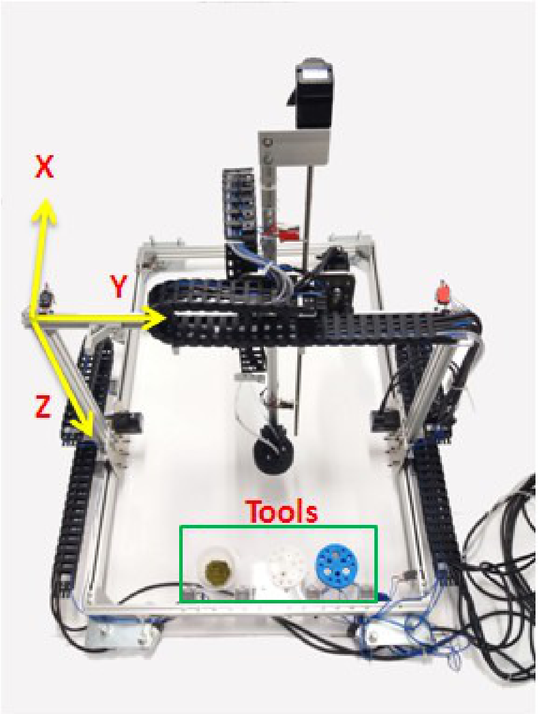


Figure 1. Cartesian farming robot prototype

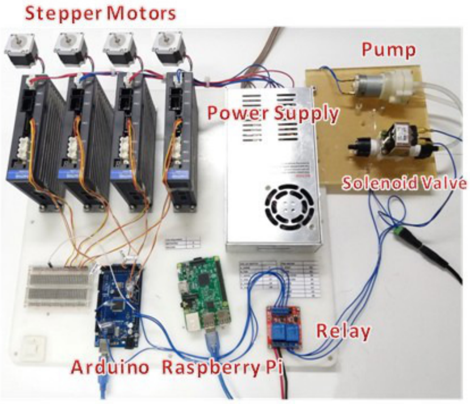


Figure 2. Control system hardware for FarmBot

For the lower level control, an arduino board is installed with the FarmBot using open-source arduino firmware [5]. This is written in C++ language and use to control the motors for movement and the tools for crop-growing functions.

As shown in Figure 2, an Arduino board Mega 2560 is responsible for the simultaneous control of all four linear bipolar motor drivers, thus the universal tool mount can be moved to any desired location with any tools including the watering tool and the seeder. In addition, the Arduino board is used to control two normally close solenoid valves for the watering tool and a vacuum pump for the seeder.

Figure 3, show the arduino receives the commands from the PC application in the form of G-code through a USB cable from a Raspberry Pi controller. The commands were sent through the socket programing [6]. Some commands can be automatically modified by the decision support system before being executed.

3. Software for FarmBot

In our FarmBot program, we make a connection between PC and FarmBot using socket programming in order to create a communication between them. This includes a server and a client as its two main sections where the server listens to the specific port and IP to which clients are connected.

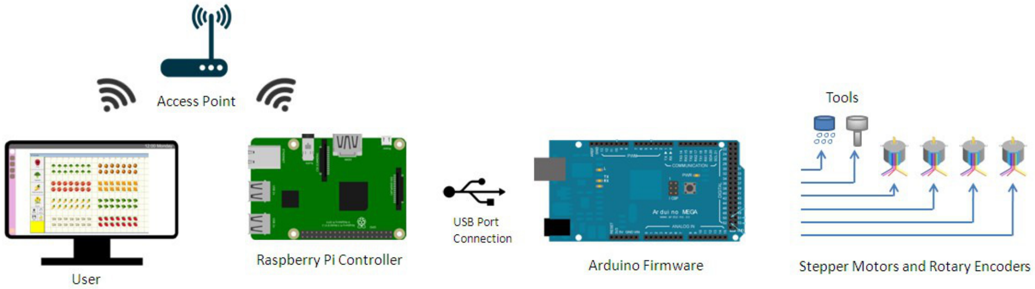


Figure 3. FarmBot communication system

Raspberry-pi (as server):

In our work, the Raspberry-pi is working as a server. Clients can connect to IP of this Raspberry-pi in order to communicate. The raspberry-pi sends commands to the Arduino which is responsible for motions of the FarmBot's motors.

Firstly, the program creates a connection between Raspberry-pi and Arduino where Raspberry-pi and Arduino communicates through serial-communication. Therefore, it adds a python serial library first and next defines Arduino's port and makes a connection.

Secondly, the Raspberry-pi is defined as a server with its IPv4 address. Then, the program has to bind this server to its IP address and port 8000. The server can listen to a number of clients at a time. Thus, the server is ready to accept commands from client. Then, using a loop we make our server to receive commands from clients iteratively otherwise the server can take only one command and get close. Noted it is using the "UTF-8" encoding format for inputs.

Client:

Any client can connect to this server using the IP address and the port number. Once connection is established, client is ready to send the signals to the Raspberry-pi over the local network.

GUI (Graphical User Interface):

To control FarmBot, the program needs a GUI system which is known as graphical user interface. We used python coding language, since a huge number of GUI frameworks/toolkits are available in it. For example, TkInter [5] is GUI programming toolkit. This is most commonly used one to create GUI applications. As Shown in Figure 5, our GUI has a base window which refers to our virtual-farm and this window can be resized in the ratio of actual farm's length and width and grid lines on it gives an idea of actual length (C).

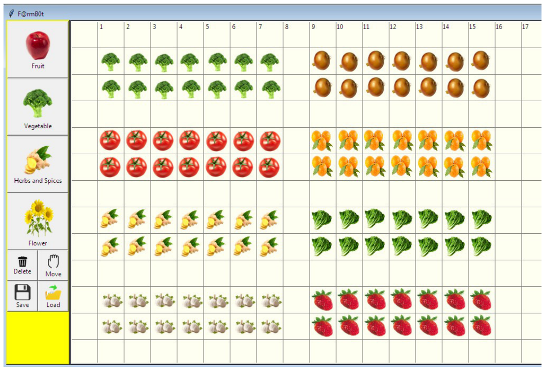
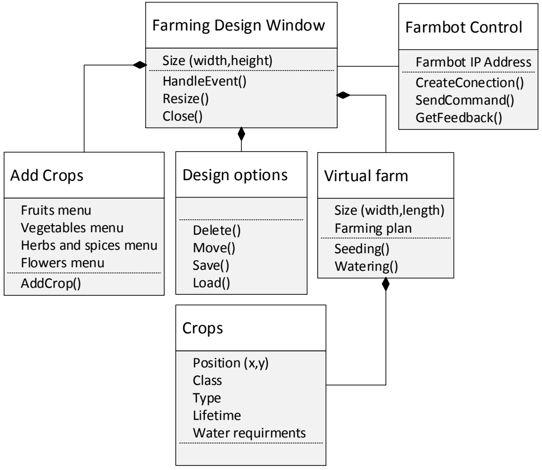


Figure 5. GUI of FarmBot window

As given in Figure 5, the menu bar (A1) on the left side of the window gives different classes of crops, fruits, vegetables, flowers, herbs and spices. User can choose any class based on his desire by just clicking on the button with the images of all famous crops. (A2) Using our application, the user can select the all major fruits, vegetables, flowers, herbs and spices which can be planted in the farm.

Further, our software application provides the options of delete, move, save and load to customize farming plan (B). Therfore, by using the given crops information and the commands the design of the farming plan and the control of the FarmBot can happen automatically. The overall process of Farmbot program is show in the class diagram as Figure 6.


Farming Design Window

Size (width,height)

HandleEvent()
Resize()
Close()

Farmbot Control

Farmbot IP Address

CreateConection()
SendCommand()
GetFeedback()

Add Crops

Fruits menu
Vegetables menu
Herbs and spices menu
Flowers menu

AddCrop()

Design options

Delete()
Move()
Save()
Load()

Virtual farm

Size (width,length)
Farming plan

Seeding()
Watering()

Crops

Position (x,y)
Class
Type
Lifetime
Water requirments

Figure 6. The class diagram of Farmbot program

4. Results and Discussion

The comparison of FarmBot web application and developed FarmBot program is given in Table 1. Our FarmBot program can work in local network in a situation where the internet connection is not available.

Table 1. Comparison of FarmBot's software

	Open source web application	**Developed FarmBot program**
Connection	Need internet connection	Run in local network
Control	Have some latency to control	FarmBot run commands suddenly
Development	Difficult to understand source code for modifying	Convenient to develop additional function in future

Further, it can run commands quickly since the commands are sent directly from the PC. Moreover, the implemented FarmBot software can be developed to convey the application and future scope for automatic agriculture systems such as new crops for seeding, new tools for planting, new sensors for measuring some parameters and etc.

5. Conclusion

In this paper, a software for an automatic farming Cartesian robot as known as FarmBot has been successfully implemented. Our software consists of two main parts, namely communication and GUI control system. For communication part, we use socket programming to build the connection between user PC and the FarmBot. It is convenient for a user to design farming plan with GUI from our implemented FarmBot program and while it can control the FarmBot automatically in an offline situation too.

As future work, we would like to apply this software platforms for other automatic agricultural systems which can be appropriate with user's farming area. Moreover, there are many features to improve and upgrade in the FarmBot system i.e., smart sensor, image processing and farming data visualization.

References

[1] Smart farming: Hope for the world, opportunity for Thailand [Online], [Accessed: August 2018], Available: https://www.scbeic.com/en/detail/product/2812.

[2] FarmBot Official Website [Online], 2014 [Accessed: August 2018], Available: https://farm.bot/.

[3] Y. Monplub, N. Intaratanoo, C Wichitphan, P. Sasithong, G Phanomchoeng and L. Wuttisittikulkij, "A Prototype of Automatic Cartesian Farming Robot," *Proceedings of ITC-CSCC 2018 Conference* (pp. 482-485), Bangkok, Thailand.

[4] FarmBot web application [Online], 2014 [Accessed: August 2018], Available: https://FarmBot.io/.

[5] FarmBot Arduino controller [Online], 2014 [Accessed: August 2018], Available: https://github.com/FarmBot/FarmBot-arduino-firmware/.

[6] M. Xue and C. Zhu, "The Socket Programming and Software Design for Communication Based on Client/Server," *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, Chengdu, 2009, pp. 775-777.

A 3D Printing with Feature-based Extension

Nopdanai Ajavakom[1], Nipit Congpuong[2], Peeratach Ritthikarn[3], Ratchatin Chancharoen[1]
[1]*Department of Mechanical Engineering, Faculty of Engineering, Chulalongkorn University*
254 Phaya Thai Road, Wang Mai, Pathum Wan, Bangkok 10330, Thailand
Ratchatin.c@chula.ac.th

Abstract

The NC programming language is developed for effective information flow in CAD-CAM-CNC chain. The international standards for the NC languages are G-code (ISO 6983) and STEP-NC (ISO 14649). Recently, additive manufacturing (AM) becomes challenging fabrication process while the computing power of a microprocessor is dramatically enhanced. Thus, the higher-level NC language that is portable and interoperable is demanded. The feature-based extension is proposed with demonstration of its benefits. The proposed NC language interfaces CAD-CAM-CNC with a measurement device, a Camera in this case. With the proposed NC language and the mockup intelligent NC machine, a 3D cylindrical feature that its position and size are to fit the existing objects are successfully 3D printed.

Keywords-G-code, STEP-NC, 3D printer

1. Introduction

G-code, developed in the 1960s and set as an ISO standard (ISO 6983) in 1970s, is an industrial standard language to effectively interface design with manufacturing so that parts are manufactured to create precision physical models. It is also the basic language within most of Computer Numerical Control (CNC) Machines today[1]. Computer-aided manufacturing (CAM) is a programming tool that can be used to generate a Numerical Control (NC) program, normally written in G-code, that is used to operate the CNC machine. The G-code language is distinctive in machine control since it effectively defines the motion of a machine and handle coordinates during a manufacturing process. Basically, there are four types of motion in G-code: 1) rapid motion: G00 2) linear motion: G01 3) counterclockwise circular Motion: G02 and 4) clockwise circular motion: G03.

Robots are also commanded in motion language, similar to G-code, e.g., G00 is similar to Point to Point (PTP) and G01 is similar to linear interpolation in robotics. However, there is no industrial standard for robots as most vendors develop their own proprietary language. However, most robotic languages include basic motion commands that are compatible with G00, G01, G02, and G03. The difference is that travel speed in a standard unit can be defined in G-code but accuracy can be defined in the robotic language. In other words, accuracy is not well defined in G-code while travel speed in the standard unit is not well controlled in the robotics system. The robotic languages include practical computer programing in their languages such as variable, branching, time, I/O and communication.

3D printer and additive manufacturing (AM) are also used G-code to operate their machines during its fabrication process. However, additive manufacturing is more complex than subtractive manufacturing, and the supported technology is far more advanced at the age of this AM technology. Thus, RepRap G-code has improved ISO 6983 so that it is effectively capable of planning, scheduling and controlling its additive manufacturing process.

Roughly, G-code is excellent at defining and controlling the motion of a machine. However, G-code focuses on the motion of the machine with respect to the machine axes, not the task [2, 3]. The G-code program is thus machine specific and cannot be used to control the other machine directly. The G-code programs must be generated along with information including machine brand, brand's model, and detail description of tool and peripherals [1]. Most CAM software includes a tool, called "Postprocessor" that can generate G-code program for a specific machine, tool, and cutting tool. The information flow during the integrated process is one way as CAD-CAM-CNC chain.

The other concern is that the modern CNC machines normally come with extensions [6], which are upgradable, beyond the scope of the ISO 6983 G-code to enhance the manufacturing capability. The machine program that is G-code free, portable and interoperable is demanded [4] to overcome this limitation and free the machine to new possibilities.

ISO 14649, recognized as STEP-NC, is initially developed in 1999 so that machine program is portable by using the feature-based concept. This is more like a higher-level language that commands the machine processes rather than the motion of machine tools. This means that a machine may complete the process in different ways, with the recommended manufacturing strategy, to fabricate the same model. In this concept, new features and techniques can be implemented and the resulting performance may be different. Pieces of research to implement STEP-NC machines are proposed [4]. However, STEP-NC is not widely used in modern CNC machines.

3D printer and additive manufacturing are now using the STL file format that stores the geometry of a 3D model, to interchange the 3D model with a machine. There is a 3D slicer or similar program, inside the 3D printer working system, to automatically convert the model in STL format into the RepRap G-code during the process. In this way, the interchange file is portable and interoperable and can be used with different machines, and thus becomes commercial standard today.

In this project, a feature-based fabrication and the data flow concept in a CAD-CAM-CNC chain are proposed such that a solid feature can be 3D printed on top of the pre-fabrication object [5]. The higher-level language and the data flow between design, manufacturing, and in-process measurement is proposed to enhance the total integrated process. In this way, the fabrication command is generated not only from CAD but also from an in-process measurement of a pre-fabrication or existing object. The proposed machine understands both the legacy G-code and feature-based commands. In the experiment, silicone floor and cylindrical wall are printed on top of the tile that its pose is determined using the camera within the automated process. The proposed command language is portable and interoperable while the machine is upgradable and flexible. The machine control system is a higher intelligence that translates this command and control all its capability to fabricate a feature.

2. Feature-based extension

In this work, an indirect STEP-NC programming method, as mentioned in [4], is mocked up. The program reads a STEP-NC file and then generates G-code and sends it to the 3D printer's controller. The 3D printer's controller accepts and interprets RepRap G-code and controls all the function of the printer. Noted that the program can also read image data from vision sensor and process it with the proposed feature-based extension to generate interactive G-code command.

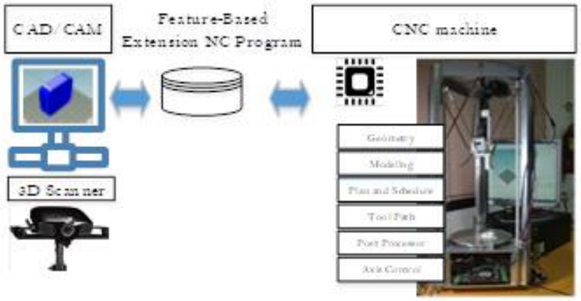


Fig 1. Feature-based NC command information flow

3. Experimental Result

3.1 A Cartesian Robot with a Silicone Head

A Cartesian robot was built with 3030 aluminum profile with feed screw and rollers. The thread of the screw is 1 mm. Each axis is driven by a two-phase stepper motor with 1/16 micro-stepping drive. This results in 1/3200 mm resolution. The workable space is 150×150×150 mm^3. In this design, the payload exceeds 2 kgs. The controller board is RUMBA 3D printer control board with Repetier firmware. The maximum travel speed is up to 50 mm/s for X and Y axes and 10 mm/s for the Z axis. The head is equipped with a silicone nozzle, Logitech camera, and laser depth sensor.

The Matlab is used as a process control program that reads the proposed feature-based NC file and then generates the RepRap G-code and sends it to the 3D printer control board. In this way, we can enjoy the programming capability of Matlab, precision real-time control of the machine, and open hardware architecture to control a fabrication process.

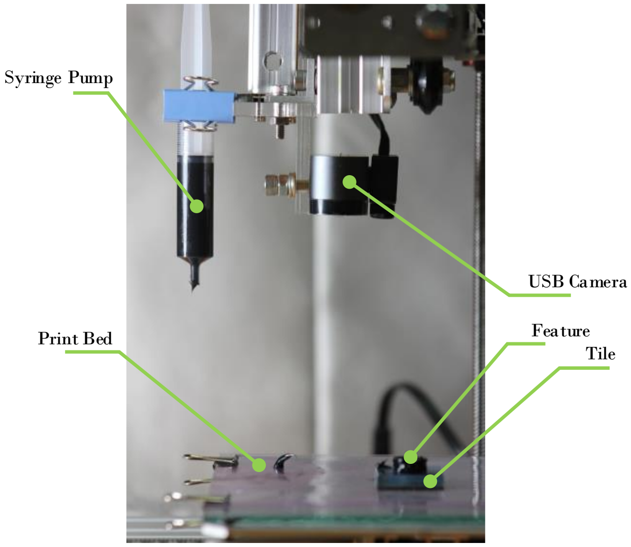


Fig 2. A Cartesian robot with a Silicone Print System

3.2 Pen Marker

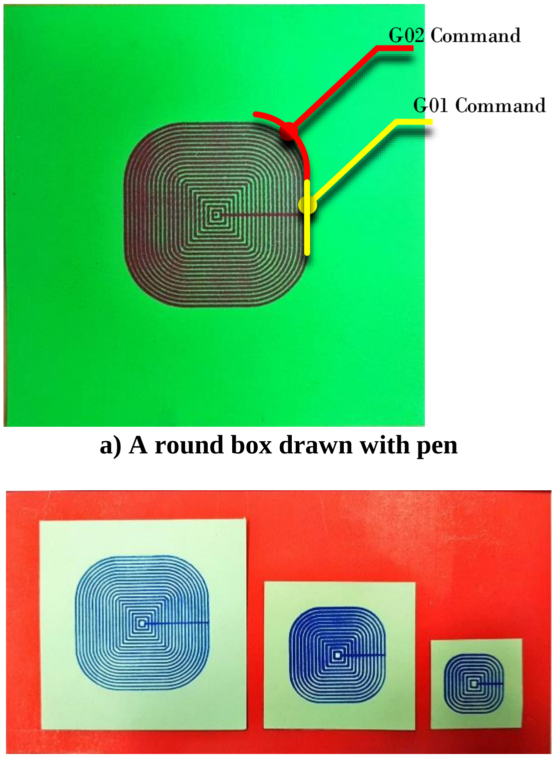
The fiber-tip pen is used in a testing process. Using fiber-tip pen can reflex some performance of 3D printer as it needs a high level of accuracy to set the specific height of pen while moving it along the path. The best solution of the testing fiber-tip pen is to set the height that the tip of the pen just touches the base of the printer. If the position of the pen is too low, the pen will contact the base too much, causing friction and unsteady line. On the other hand, if the pen's position is too high, the tip will not touch the ground and thus doesn't draw a line. The same principle applies to 3D printing nozzle as too much contact will crash the former layer, while too high tip will cause discontinuity or distorted shape.

This experiment uses probing technique to set pen's height that touches the base exactly. Probing has several advantages in advance 3D printing. It can be used to measure the base's height at any time. In this experiment that base is not a single layer plan, so the height of the second layer, or wall, can be measured before planning any printing.

Another importance of probing is to use in level revering process. In the general case, the base of the printer is not always perpendicular to the extruder. Probing can measure the difference in height and then adjust the extruder to perpendicular to the base and vice versa. If the printing continues with slant base, the might also have some distortion in printing product. In pen testing, the researchers operate the printer to probe and fill-in the square shape with circular edges on green paper with the distance between each line is 3.5 millimeter. The result is shown in Fig 5a.

Furthermore, the 3D printer can operate any available G-code command. For instance, as shown in Fig 5a, G02/G03 command, arc move, is used for making every corner of a curve edge square. Generally, the slicer program generates G01 command, linear move, at any kinds of shape because some firmware does not support G02/G03 command. However, feature-based extension NC command does not specify the movement of the 3D printer. On the other hand, when the feature-based extension commands are sent, the intelligent printer chooses and generates available and suitable movements itself. In this experiment, Matlab generates G03 at the corners. Moreover, because of this capability, the feature-based extension NC command would support any feature that will be implemented in the future.

The other capability of the feature-based extension NC command is that the printer can create the feature with real-time measurement, which is helpful for users especially when a feature is printed on top of a complicated-shape object. In this experiment, the length of the different size of color paper is measured using a USB camera and image processing method. Then, these measurements are used for calculation through Matlab and then a curved edge square is printed on the paper with the relative dimension.



a) A round box drawn with pen

b) Interactive boxes that their sizes fit the paper

Fig 3. Pen Draw with the Proposed Technique

3.3 A Silicone 3D Printing

A silicone is used as a material for demonstrating 3D print in this experiment. To generate a perfect result, many measurement and calculation have been executed. For determining the centroid and dimension of tile, the USB camera is used for taking pictures of the tile. Matlab is used for detecting image's dimension and centroid in the pixel through image processing, using built-in color thresholder function and for finding machine reference position through mapping.

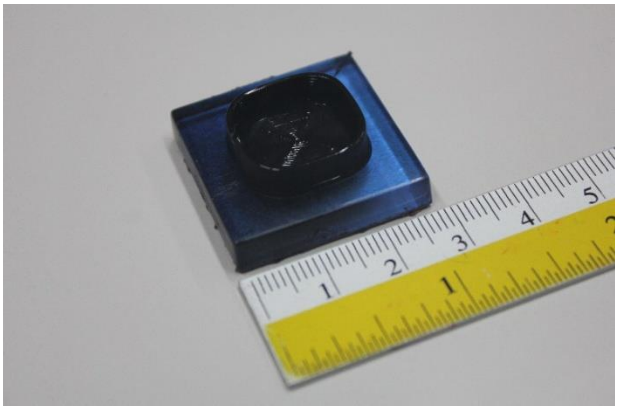
Like the pen maker, the tip height is effect on the quality of printing. The printer requires the high level of accuracy to set the specific height of the nozzle, especially for the first layer. The nozzle would drag flooded silicone if the head is set too low. In the worst case, the nozzle might hit the bed can cause damage. On the other hands, the first layer would not form if the nozzle is set too high.

As a result, the USB camera is used for determining the relative depth where the nozzle almost touches the tile since a silicone needsroom for forming the first layer. During the finding depth process, the detecting dimension process is repeated at the different height to find, on Matlab, an equation representing a relationship between the tile's dimensions in pixel and the z-axis values in millimeters.

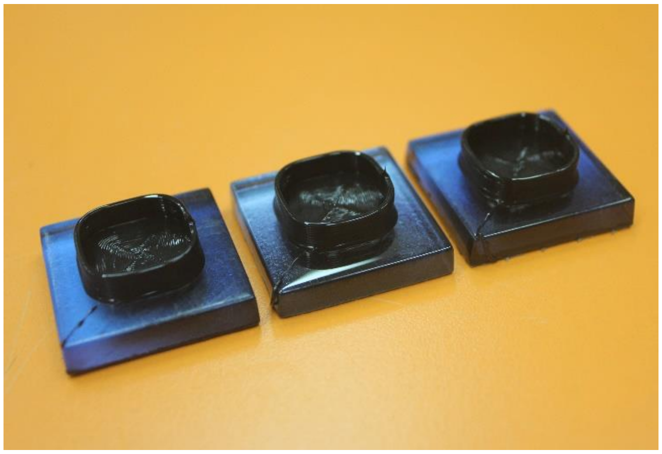
For other layers, the high level of accuracy setting is needed as well, but the solution is simple as the height between layers is equal to the diameter of the nozzle, so silicone has enough room for laying on the former layer.

For the printing process, the ratio of the area of a syringe pump to that of the nozzle is determined so that an appropriate amount of silicone is extruded between the starting point and ending point. The printing process is divided into two parts. The first part is printing the based or the first layer which is like the printing process of the pen maker, but the syringe pump is used instead of fiber-tip pen and the 3D printer's controller also controls the extruder so that the silicone is pushed throughout the printing process. The second part is printing the wall, which only the most outer boundary of the feature is printed.

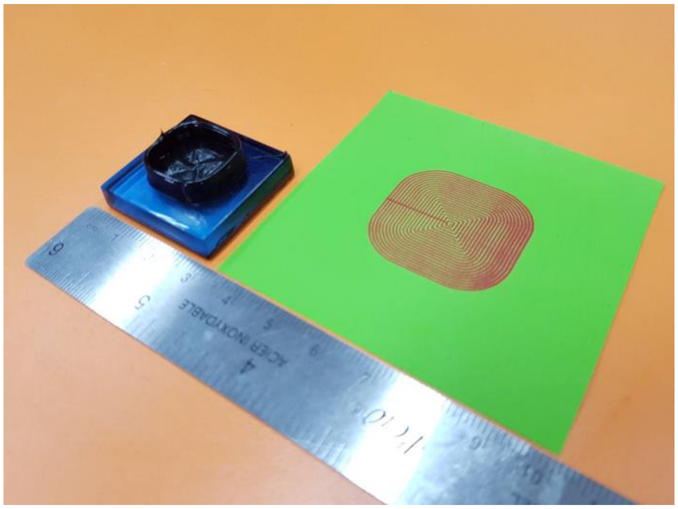
As shown in Fig. 6c, the command used for printing pen maker result is as same as that used for printing the base or the first layer of silicone 3D print model even though the print head tool is changed from fiber-tip pen to syringe pump.



a. A Silicone Feature 3D Printed on top of a tile



b. Some Silicone Results



c. Interactive Print

Fig 4. Print Results

This result demonstrates one capability of feature-based extension NC command. While G-code program, generated by slicer program, is specific to the machine, meaning that it needs to be regenerated every time after the 3D printer has been changed or modified, feature-based NC command is flexible so that the printer successfully works with the same command despite different tools such as a print head.

3.4 The proposed NC program

The proposed NC program is the G-code program with feature-based extension.

```
01 G28 ; HOME
02 VAR POS#01 // INITIALIZE VARIABLE
03 TAKE PICTURE
04 FEATURE EXTRACTION('VAR', POS#01)
05 PROBE('VAR', POS#01, 'PATTERN', 'CORNER')
06 FLOOR('VAR', POS#01, 'PATTERN', 'Default')
07 WALL('VAR', POS#01, 'HEIGHT', 20)
08 G28
```

The example reveals the proposed NC program that is used in the example experiment. The legacy G-code command is still valid. The first line in the example is G28 to command the machine to go to its home position. It is noted that the home position of each machine is different. The second line, 'VAR POS#01', is to define the variable or memory that can be utilized in the later command. The type of this variable is the homogeneous transform [10] or POSITION which defines position and orientation of a point. This type of variable is the basic type in robotics application since frame concept is often used. The 'TAKE PICTURE' command demonstrates that the modern machine is capable of interfacing withnot only simple I/O, but also an intelligence sensor or actuator, the Logitech camera in this case. The NC program should be able to cope with this. The 'FEATURE EXTRACTION' command demonstrates that not only simple arithmetic, but also complex analysis is preferred in the highly complex process. The 'PROBE' command is the higher-level motion command that will determine the bed or part's positions (Z) at the given (X, Y) positions where 'CORNER'(s) are desired in the example case. This is like "manufacturing strategy" in the STEP-NC program. This command will coordinate the machine motion to determine the result and thus it is the feature-based command. This higher level 'PROBE' command may issue some RepRap G-code G30 commands which is the probe command at a current (X, Y) position. The 'FLOOR' and 'WALL' commands are the extensions of the G-code that are in features-based command.In this investigation, the 3D printer's control board is loaded with the latest '*Repetier firmware*' where G02 and G03 are available, thus the process control program translates the NC program into the G-code low-level program that includes these circular motion commands. The looping and timing programming concepts are implemented within the process control program.

This example demonstrates that the intelligent CNC machine is essential for a higher-level language, but new capability and methodology can be utilized. This machine accepts both the legacy G-code (and M-code as well) and portable feature-based commands. To interpret and process these commands require most of the programming capabilities as stated earlier and thus computing power embedded within the CNC machine. The proposed NC language and programming power can free the NC machine to new performance and capability.

4. CONCLUSIONS

A 3D cylindrical feature that its position and size are to fit the existing object is successfully 3D printed. In the proposed workflow, the data flow in the CAD-Measurement-CAM-CNC chain and the NC language are developed. The portable and interoperable NC language demands a higher intelligent NC controller but leads to a new technique of how we build a 3D object. The feature-based NC language will free a machine for new fabrication techniques.

Acknowledgement

This work was a part of the intelligent 3D printing for development project which was supported by the Development into the 2nd Century Fund, Chulalongkorn University.

5. REFERENCES

[1] M. Minhat, V. Vyatkin, X. Xu, S. Wong, Z. Al-Bayaa, *A novel open CNC architecture based on STEP-NC data model and IEC 61499 function blocks*, Robotics and Computer-Integrated Manufacturing, Volume 25, Issue 3, 2009, Pages 560-569, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2008.03.021.

[2] X.W. Xu, S.T. Newman, *Making CNC machine tools more open, interoperable and intelligent—a review of the technologies*, Computers in Industry, Volume 57, Issue 2, 2006, Pages 141-152, ISSN 0166-3615, https://doi.org/10.1016/j.compind.2005.06.002.

[3] Xu, X.W. , Realization of STEP-NC enabled machining, (2006) Robotics and Computer-Integrated Manufacturing, 22 (2), pp. 144-153.

[4] Matthieu Rauch, Raphael Laguionie, Jean-Yves Hascoet, Suk-Hwan Suh, *An advanced STEP-NC controller for intelligent machining processes*, Robotics and Computer-Integrated Manufacturing, Volume 28, Issue 3, 2012, Pages 375-384, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2011.11.001.

[5] Patharawut Suphama, Kuntinee Maneeratana and Ratchatin Chancharoen, 2017. Positioning of Fused Deposition Features on Primitives. Journal of Engineering and Applied Sciences, 12: 3818-3823. DOI: 10.3923/jeasci.2017.3818.3823.

Cloud and Distributed Computing

Efficient Checkpoint Interval for Speculative Execution in MapReduce

Naychi Nway Nway
University of Information Technology
Yangon, Myanmar
naychinwaynway@uit.edu.mm

Ei Chaw Htoon
University of Information Technology
Yangon, Myanmar
eichawhtoon@uit.edu.mm

Abstract

The MapReduce has become popular in big data environment due to its efficient parallel processing. However, MapReduce still has the problem from job delay caused by straggling tasks, which prolong job completion time. In MapReduce framework, although the existing speculative execution mechanism mitigate stragglers, its tasks are slower than their original tasks so this makes job completion time get long when straggling tasks occur. So, in this paper, a checkpoint mechanism is proposed in order to increase the efficiency of speculative execution of MapReduce, and not to prolong job completion time in case of straggling tasks. However, MapReduce produces too much intermediate data; as a result, checkpoint of every intermediate data can still decrease the performance of MapReduce. So, to avoid this problem, the proposed system evaluates checkpoint interval in order to reduce job completion time in case of stragglers. Then, the proposed system defines stragglers using LATE scheduler. The proposed checkpoint interval is based on five parameters: expected job completion time without checkpointing, checkpoint overhead time, rework time, down time and restart time. Experimental results show that the proposed system leads to less completion time, rework time and checkpoint overhead.

Keywords- MapReduce, straggling task, big data, checkpoint interval, completion time

1. Introduction

Data-intensive applications process vast amounts of data with special-purpose programs. Even though the computations behind these applications are conceptually simple, the size of input datasets requires them to be run over thousands of computing nodes [6]. For this, Google developed the MapReduce framework [5], which allows non-expert users to run complex tasks easily over very large datasets on large clusters. The large datasets are often messy that causes I/O overload and contain skewed data. This may, in turn, cause a task or even an application to be long completion time. It points out that MapReduce has a performance problem while slow tasks also called stragglers occur.

The impact of stragglers can be considerable in terms of performance. In MapReduce process, after map stages, the intermediate data is produced and it is the input for reduce stages [1]. So, intermediate data is important to be a successful MapReduce process. Although MapReduce can restart the process and produce intermediate data again when slow tasks occur, it can prolong job completion time.

A few of straggler mitigation techniques have been developed and can be divided into two classes: black-listing and speculative execution [9]. Blacklisting uses a user-provided health-check script to detect the status of the slaves. If a slave is not performing properly, it can be blacklisted so that no job will be scheduled to run on it. However, a strict or incorrect health-check program will result in reduced numbers of resources. Besides, stragglers can arise on the non-blacklisted machines at times, often due to some complex reasons like I/O contentions, background services, and hardware behaviors. In speculative execution, the master schedules speculative tasks for those straggling tasks and puts them in the queue. They will be launched when there are available slots. For each original task, the scheduler also ensures that at most one speculative task is running at a time. The original task is killed if the speculative task finishes first and vice versa.

Although the original speculative execution has fault-tolerance feature, it has drawback because of re-executing tasks from start as their original tasks. It re-reads the input data, re-copies the intermediate data and re-computes the processed data so straggling tasks cause the job completion time to take longer [9].

Therefore, in this paper, checkpoint interval-based speculative execution is proposed to reduce the job completion time when straggling tasks occur in Hadoop MapReduce. This proposed checkpoint interval is calculated before starting the process of map tasks. After defining checkpoint interval, checkpoint file is created in local disk of a node and takes checkpoint according to proposed checkpoint interval. The proposed system evaluates the performance of job completion time based on mean time between slow tasks, which is the expected time between two slow tasks for a repairable system. The evaluations measure the performance of job completion time of the proposed system, original MapReduce and one of the related works. And then, the experiments show that this proposed system takes less overhead, completion time and rework time because of proposed checkpointing strategy.

The paper is structured as follows:

the related work of proposed system is discussed in Section 2. Section 3 explains the basic flow and built-in speculative execution of MapReduce. The checkpoint interval and implementation of proposed system are described in Section 4. Section 5 describes the experimental results and finally, the conclusion of this paper is presented in Section 6.

2. Related Work

MapReduce [1] is a parallel programming model which is originally proposed by Google in 2004 to deal with the rapidly increasing demand of processing mass data concurrently. Through well-defined interfaces and runtime support library, MapReduce can automatically perform the large-scale computing tasks in parallel, hide the underlying implementation details, and reduce the difficulty of parallel programming, which makes MapReduce become one of the most widely used parallel programming models in the concurrent processing vast amount of data.

RAFTing MapReduce presented in [6] tries to create several kinds of checkpoint to handle different failures. RAFT-LC is a local checkpointing algorithm that allows a map task to store progress metadata on local disk and later restores based on this in case of failures. RAFTing mappers push data to reducers instead of the opposite way and make the intermediate data replicated without bringing much overhead.

In [7], authors also proposed new scheduling algorithm in order to improve the speculative re-execution of straggling tasks in MapReduce. ESMAR differentiates historical stage weight information on each node and divides them into k clusters in order to identify straggling tasks accurately.

In paper [9], the author introduced two checkpoint algorithms to eliminate the costs of re-reading, re-copying, and re-computing the partially processed data. It makes an input checkpoint to record the location of unprocessed input data, while the output checkpoint consists of spilled files and their index information. Yong proposed a first-order model that defines the optimal checkpoint interval in terms of checkpoint overhead and mean time to interrupt (MTTI). Yong's model does not consider failures during checkpointing and recovery [8].

Given the checkpointing parameters such as checkpoint latency and MTTI, Daly's model [3] provides a method for computing the optimal checkpoint which is associated with the optimal execution time. Checkpoints are created when the progress reaches 0.5 (or) 0.25 by calculation progress rate and estimated task execution time [2].

In original version of MapReduce [1], all of the straggling tasks are re-executed again in case of slow tasks. As a result, the job completion time can be long because of starting the tasks from scratch. In work [2], when the checkpoints are created in 25% of execution time, the speculative execution before 25% is not recovered. To overcome the problem of previous work in [1] and [2], the proposed system defines a checkpoint interval that influences the number of checkpoint operations performed during an application's execution. To ensure that checkpoints can be used effectively, the proposed system evaluates checkpoint interval and finds stragglers using LATE scheduler that aims to recover from straggling tasks and to improve performance as the main goal. Unlike original MapReduce, the proposed system reschedules the straggling tasks without starting again. The experiments show the performance comparison among original MapReduce, the proposed system and one of the related work [2].

3. The MapReduce Framework

3.1. Execution Flow of MapReduce

MapReduce [4] adopts a two-stage and shared-nothing design. The first stage, the map stage, takes a list of key value pairs as input, and applies a map function on each of the pairs to generate arbitrary number of intermediate key value pairs. In the second stage, all the intermediate values associated with the same keys are grouped together as a list, and a reduce function takes each of the groups as input to generate another arbitrary number of final output key value pairs. The paradigm behind MapReduce is a quite simple behavior because a map or reduce function calls on a key value pair that shall depend neither on other pairs nor on the processing order. This makes it easy to split the whole job into smaller independent subtasks that can run in parallel.

The input data files of MapReduce are usually stored on a DFS (distributed file system) such as HDFS, an open-source implementation of GFS. The data files are split into small pieces logically, every one of which will be fed to a map task. Map tasks, also known as mappers, parse raw input data that splits into k1 v1 pairs, and invoke the map function on every single pair, the generated k2 and v2 pairs are written to a memory buffer. When the buffer verges to overflow, the mapper flushes it to a local disk file, which is called a spill. A mapper may create several spill files, however, it will merge the spill files into a single output file on local disk after all input records are processed.

There are usually several reduce tasks, or reducers, key value pairs with the same key hash value that goes to the same reducer. As a result, the single map output file shall be logically spilt into parts; each part will be fed to a reducer. A reduce task can be summarized to 3 main phases: shuffle, sort and reduce. During the shuffle phase, reducers copy outputs from each mapper, and merge the outputs into fewer amounts of files in the sort phase. The shuffle phase and sort phase often overlap in practice, but the reduce phase shall not start until the shuffle phase finishes, which is limited by the MapReduce semantics.

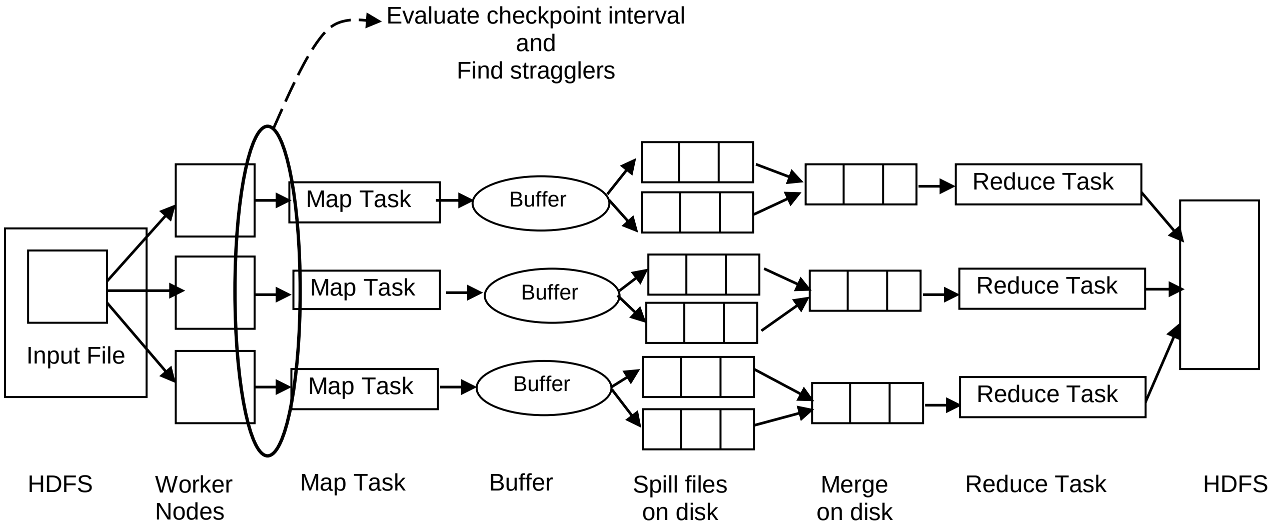


Figure 1. Proposed system architecture

3.2. Speculative Execution in MapReduce

Firstly, all the tasks for the jobs are launched in Hadoop MapReduce. The JobTracker monitors the progress of each task using a progress score between 0 and 1. The average progress score of each category of tasks (maps or reduces) is used as the threshold for speculative execution: if a task's progress score is less than the average minus 0.2, it is considered as a straggler. The speculative tasks are launched for those tasks that have been running for some time (at least one minute) and have not made any much progress, on average, as compared with other tasks from the job. The speculative task is killed if the original task completes before the speculative task, on the other hand, the original task is killed if the speculative task finishes before it [11]. However, speculative execution re-executes from start as their original tasks so speculative execution in MapReduce cause the job completion time to get long although it has fault-tolerance features. So, this paper uses LATE[7] which defines a task is straggling or not. After that, based on expected job completion time, the formulated checkpoint interval is proposed in order to keep going after straggler tasks. So, the proposed system can save a lot of time when straggling tasks are involved.

4. Proposed System

In this paper, a checkpointing strategy for MapReduce is proposed, which defines checkpoint interval to improve the efficiency of checkpoint in speculative execution and the job completion time. To preserve this, Figure 1 shows the architecture of the proposed system.

4.1. Expected Job Completion Time without Failure

Although original MapReduce processes with its own speculative execution for stragglers, it reworks a task from start. So, stragglers in MapReduce make a job completion time long because they require finished process ranges to be executed again. The main design goal of this proposed system is to provide a checkpointing strategy by permitting the tasks to checkpoint at formulated checkpoint interval.

Initially, the input file is taken from HDFS and InputFormat class is used to split the input into multiple file splits. After dividing the file, this proposed system will calculate checkpoint interval, and then, based on this interval, creates the checkpoint to keep track of progress of MapReduce job. All of task progresses are saved in checkpoint file before the execution of one Mapper task. The checkpoint file is saved in local disk of the node that runs the current MapReduce process so the node can restart tasks from recent status with the help of checkpoint file when straggling tasks occur. To calculate the proposed checkpoint interval, firstly, the system calculates the expected job completion time [4] without checkpoint using (1)

$$T_{c=} \left(\frac{T_n}{w}\right) * \left(J_t + \frac{Dsize}{J_p}\right) \tag{1}$$

where T_c means job completion time, T_n means the numbers of tasks, w means number of workers, J_t means time to take JVM, Dsize means input data size and J_p means processing size of JVM per second.

4.2. Checkpoint Interval Model

The proposed checkpoint interval is based on Daly's model [3] except downtime parameter. The proposed system adds downtime parameter because there are many map tasks in MapReduce, which are important for successful completion of a MapReduce job. So, the downtime is needed to consider as a parameter for calculating checkpoint interval. The checkpoint interval model is defined by five parameters given in Table 1.

Table 1. Checkpoint interval parameters

Parameters	Description
M	Mean Time Between Slow Tasks
Β	Checkpoint Overhead-time to take a checkpoint file
R	Restart Time- time required before an application resumes to current work
Rework Time	Time needed to rework job due to slow tasks
D	Down Time-time that cannot arrive current running state in case of slow tasks

Based on job completion time, the system calculates interval between checkpoint files that minimizes the time lost when slow tasks occur using (2).

T = Completion Time + Overhead Time + Rework Time+Down Time+Restart Time (2)

Completion Time is defined as actual completion time without checkpoints. Completion Time will be Tc and Overhead Time will be β(C(τ)-1) where C(τ) is the number of checkpoint taken and one is subtracted because there is no need to write checkpoint files in the last segment. For Rework Time, it will be described by ½(τ+β)N(τ) where N(τ) is the expected numbers of interrupt. Down Time is used as DN(τ) and finally, Restart Time is RN(τ), the amount of time required to restart into total number of slow tasks. So, the system constructs the formula as (3)

$$\mathrm{T} = T_c + (C(\tau) - 1)\beta + \frac{1}{2}(\tau + \beta)N(\tau) + DN(\tau) + RN(\tau) \tag{3}$$

Next, the system determines the numbers of interrupt N(τ)and numbers of checkpoints are calculated by dividing completion time by checkpoint interval. The expected numbers of interrupt can be calculated by the product of numbers of checkpoints required to complete calculation and the probability of each segment failing as in (4)

$$N(\tau) = \frac{Tc}{\tau}\left(e^{\frac{\tau+\beta}{M}} - 1\right) \cong \frac{Tc}{\tau}\left(\frac{\tau+\beta}{M}\right) \tag{4}$$

Then, N(τ) is substituted in (3):

$$T = Tc + \left(\frac{Tc}{\tau} - 1\right)\beta + \left[\frac{1}{2}(\tau + \beta) + D + R\right]\frac{Tc}{\tau}\left(\frac{\tau + \beta}{M}\right) \tag{5}$$

Using (5), the system finds the minima with respect to τ that set the derivation to zero.

$$e^{\frac{\tau+\beta}{M}}[\tau^2 + (\beta + 2R + 2D)\tau - (\beta + 2R +)M] + 2RM - \beta M = 0 \tag{6}$$

Instead of expanding the exponential term, recast (6) as follows:

$$\frac{\tau + \beta}{M} = ln\left[\frac{(\beta - 2R)M}{\tau^2 + (\beta + 2R + 2D)\tau - (\beta + 2R + 2D)M}\right] = \ln[g(\tau)] \tag{7}$$

The system which calculates a Taylor series expansion for natural logarithm of g(τ) is as follows:

$$\frac{\tau + \beta}{M} = \frac{g(\tau) - 1}{g(\tau)} + \frac{1}{2}\left(\frac{g(\tau) - 1}{g(\tau)}\right)^2 + \frac{1}{3}\left(\frac{g(\tau) - 1}{g(\tau)}\right)^3 + \cdots$$

$$= \left(1 - \frac{1}{g(\tau)}\right) + \frac{1}{2}\left(1 - \frac{1}{g(\tau)}\right)^2 + \frac{1}{3}\left(1 - \frac{1}{g(\tau)}\right)^3 + \cdots \tag{8}$$

Reduce the (8) to quadratic form as in (9)

$$\tau^2 + 2D\tau + (\beta^2 - 2\beta(R + M) - 2DM) = 0 \tag{9}$$

Finally, the value of τ which minimize (5) as follows:

$$\tau = -\beta + \sqrt{2\beta(R + M) + 2DM} \tag{10}$$

According to the above derivation, checkpoint interval for MapReduce process can be calculated using (10). The input for checkpoint interval is checkpoint overhead, restart time, mean time between slow tasks and down time of a MapReduce job.

4.3. Speculative Execution in Proposed System

After evaluating checkpoint interval, the system checks stragglers using LATE scheduler. To select tasks for speculative re-execution, Hadoop default scheduler monitors the progress of tasks using Progress Score (PS) between 0 and 1. Suppose: a job has K number of tasks being executed; a task has a total of N number of key/value pairs to be processed and M of them have been processed successfully. Hadoop default scheduler gets PS according to (11).

$$PS = \left\{\begin{array}{ll} \frac{M}{N} & For\ Map\ Task \\ \frac{1}{3} * \left(K + \frac{M}{N}\right) & For\ Reduce\ Task \end{array}\right\} \tag{11}$$

$$PSavg = \sum_{i=1}^{K} PS[i]/K \qquad (12)$$

$$\text{For task } Ti\text{: } PS[i] < PSavg \text{ -20\%} \qquad (13)$$

Here, it is assumed that a map task spends negligible time in the order stage and a reduce task has finished K stages and each stage takes the same amount of time. If (13) is satisfied, task *Ti* needs a backup task. The backup task is started from the last checkpoint interval; as a result, it saves not only completion time but also rework time.

5. Experimental Results

To evaluate the effectiveness of this proposed system in the presence of straggling tasks, the mean times between slow tasks are thought of the thing. That is, defining values of mean time between slow tasks in order to consider the job completion time that is measured from performance aspect of the proposed system. Compare the checkpoint overhead aspect and rework time in the case of straggling tasks. The implementation of the proposed system is based on Hadoop 2.7.4, Java 1.8 and Hadoop Distributed File System (HDFS) with data size of 1GB. The jobs for experiments are word count over user-submitted comments on StackOverflow. The proposed jobs contain 8 map tasks and 1 reduce task, each map task processes about 128 megabytes of data.

In scenario with only slow tasks, Figure 2 shows the relationship between job completion time and numbers of checkpoint. It introduces mean time between slow tasks 20 which means slow tasks occur too frequently. This shows that when slow tasks occur frequently, the system needs to take more checkpoints in order to save completion time. To avoid making completion time long, the numbers of checkpoint should be taken carefully.

As shown in Figure 3, the comparison among the proposed system, original MapReduce and one of related works whose checkpoint intervals is 25% of execution time. In accordance with Equation 10, checkpoint intervals are calculated based on different mean time between slow tasks. According to the experiment, the proposed system takes less completion time not only in mean time between slow tasks 100 but also in mean time between slow tasks 20. As a result, the proposed checkpoint interval works efficiently in the case of slow tasks that occur in MapReduce. Although, the completion time of proposed system is slightly the same with related work, the completion time is decreased when slow tasks appear frequently.

As another comparison aspect, as in Figure 4, the experiment will show the checkpoint overhead aspect of proposed system. The values of x-axis are checkpoint intervals that are obtained by calculating Equation 10. In The checkpoint interval values are calculated based on mean time between slow tasks from 10 to 100 in seconds. Figure 4 shows the job completion time under different values of checkpoint overhead. We compare three different checkpoint overhead times, C=5, C=3 and C=1 in seconds. For these experiments, start time and down time take 2 seconds. The experiment shows that slightly difference checkpoint overhead that is negligible for our proposed system. So, our proposed system is suitable not only checkpoint overhead in 1 second but also checkpoint overhead in 5 seconds.

Figure 5 shows the performance of proposed system based on rework time. It is shown that along with straggler tasks, the proposed system significantly decreases job completion time compared with other systems because of proposed checkpoint interval. The proposed system can also save rework time because the system continues the work from last checkpoint in case of straggling tasks.

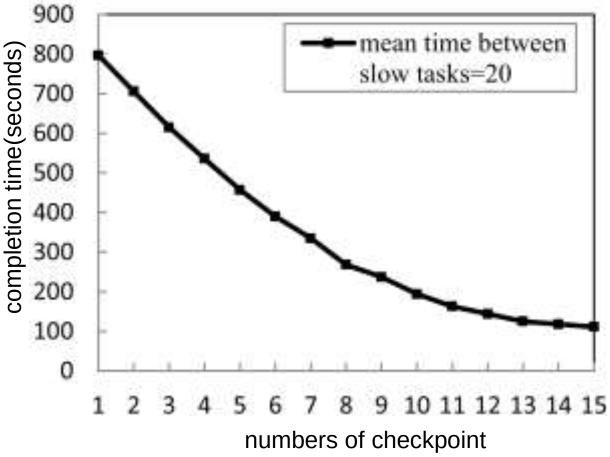


Figure 2. Job completion time versus numbers of checkpoint

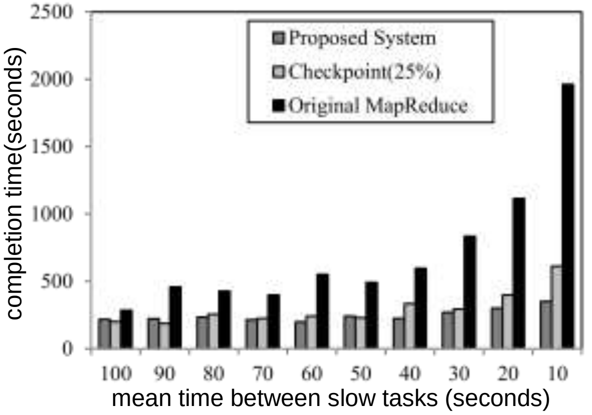


Figure 3. Comparison of completion time with checkpoint overhead=5s, restart time=2s and downtime=2s

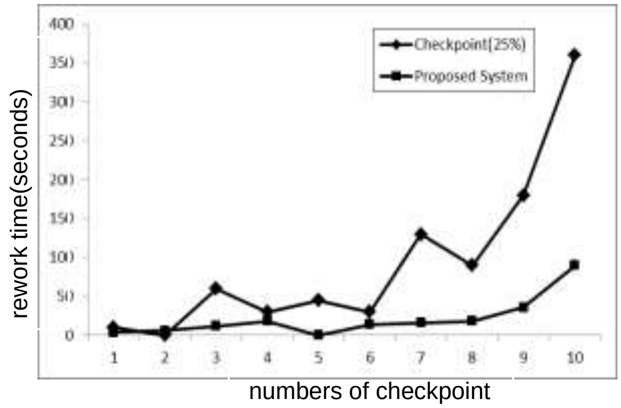


Figure 5. Comparison of completion time based on rework time

6. Conclusion

MapReduce is a popular programming model that allows the user with simple APIs and is able to run big data applications. The popularity of MapReduce is that it makes the parallelization easy and has speculative execution strategy. Although MapReduce is able to retry the straggling tasks, it performs poorly because it re-executes all finished ranges again in case of stragglers. As a result, MapReduce job can prolong job completion time when straggling tasks occur.

To overcome the limitation of existing speculative execution in MapReduce, the proposed system uses checkpointing strategy in order to avoid re-execution of finished tasks in case of straggling tasks. Proposed checkpointing mechanism which defines the most suitable interval to take checkpoints, as a result, saves job completion time, rework time and checkpoint overhead.

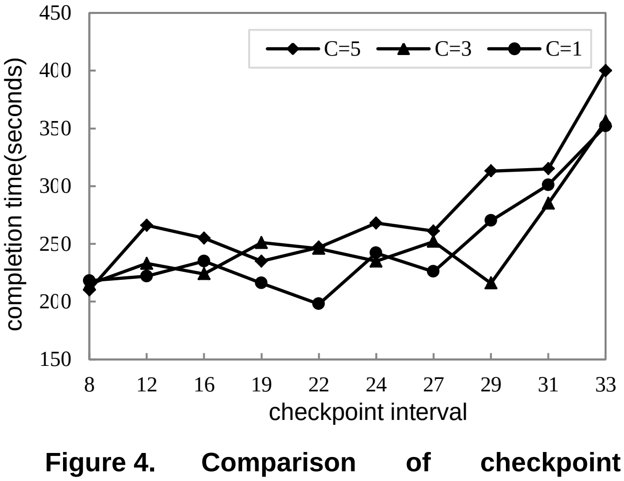


Figure 4. Comparison of checkpoint overhead

The proposed system implemented on the base of Hadoop that is the most popular open source implementation of MapReduce. The proposed system outperforms original MapReduce while decreasing mean time between slow tasks.

7. References

[1] B. Cho, I. Gupta, “Making cloud intermediate data fault-tolerant,” ACM symposium on cloud computing,2010.

[2] C. Lin, T. Chen and Y. Cheng, “On improving fault tolerance for heterogeneous Hadoop MapReduce clusters,” IEEE International Conference on Cloud Computing and Big Data, 2014.

[3] D. John, “Future generation computer systems,” vol. 22,Issue 3, February 2006, pp. 303–312.

[4] H.Wang, H.Chen and F.Hu, “BeTL: MapReduce checkpoint tactics beneath the task level,” IEEE Transactions on Services Computing,2016.

[5] J.Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” 6th symposium on operating system design and implementation (OSDI), San Francisco, December 2004.

[6] J.Quiane Ruiz, C. Pinkel, J.Schad and J. Dittrich, “RAFTing MapReduce: Fast recovery on the RAFT,” IEEE International Conference on Data Engineering,2011.

[7] L.Ying, H.Chen and S.Xiaoyu, “ESAMR: An Enhanced Self-Adaptive MapReduce Scheduling Algorithm,” IEEE 18th International Conference on Parallel and Distributed Systems, 2012.

[8] W. Yong, “A first order approximation to the optimum checkpoint interval,” ACM 1974.

[9] Y.Wang, W. Lu, R. Lou and B.Wei, “Journal of grid computing,” vol.13, Issue 4, December 2015, pp. 587–604 .

[10] Sorting 1PB with MapReduce: http://googleblog.blogspot.com/ 2008/11/sorting-1pb-with-mapreduce.html.

[11] https://data-flair.training/blogs/speculative-execution-in-hadoop-mapreduce/

A Ciphertext Policy Attributes-based Encryption Scheme with Policy Revocation

Phyo Wah Wah Myint, Swe Zin Hlaing, Ei Chaw Htoon
University of Information Technology
Yangon, Myanmar
phyowahwah@uit.edu.mm, swezin@uit.edu.mm, eichawhtoon@uit.edu.mm

Abstract

There are a lot of *data exchanges among the parties by using cloud computing. So data protection is very important in cloud security environment. Especially, data protection is needed for all organization by security services against unauthorized accesses. There are many security mechanisms for data protection. Attributes-based Encryption (ABE) is a one-to-many encryption to encrypt and decrypt data based on user attributes in which the secret key of a user and the ciphertext are dependent upon attributes. Ciphertext policy attributes-based encryption (CP-ABE), an improvement of ABE schemes performs an access control of security mechanisms for cloud storage. In this paper, sensitive parts of personal health records (PHRs) are encrypted by ABE with the help of CP-ABE. Moreover, an attributes-based policy revocation case is considered as well as user revocation and it needs to generate a new secret key. In proposed policy revocation case, PHRs owner changes attributes policy to update available user lists. A trusted authority (TA) is used to issue secret keys as a third party. This paper emphasizes on key management and it also improves attributes policy management and user revocation. Proposed scheme provides a full control on data owner as much as he changes policy. It supports a flexible policy revocation in CP-ABE and it saves time consuming by comparing with traditional CP-ABE.*

Keywords- Attributes-based encryption (ABE), Ciphertext policy attributes-based encryption (CP-ABE), Personal Health Records (PHRs), Trusted Authority (TA)

1. Introduction

Modern societies and organizations are motivated to outsource more and more sensitive information into the cloud servers. Protecting data from unauthorized users and other threats is a very important task for security providers. ABE performs as an attributes-based access control with an encryption mechanism for data confidentiality. ABE allows users to encrypt and decrypt data based on user's attributes. In ABE, if the attributes of a user satisfy an access structure of ciphertext, the user can get a secret key associated with that ciphertext. Collusion-resistance is crucial security feature of ABEs. Another modified form of ABE is Key-Policy ABE (KP-ABE) as shown in Figure 1. In KP-ABE scheme, data owner cannot decide a user who can decrypt the encrypted data. The problem is that it can only choose descriptive attributes for the data [4] [10]. Then, another modified form of ABE is CP-ABE as shown in Figure 2. CP-ABE improves the existing ABEs because the encryptor can choose the decryptor who can decrypt a cipher. It can support an access control in the real environment [4] [10]. CP-ABE has still limitations in terms of specifying policies and managing user attributes [4] [6]. In this paper, a policy revocation scheme is added in traditional CP-ABE scheme. Traditional CP-ABE scheme has not considered policy revocation case. A sample PHRs data sharing scenario is shown in Figure 3. For PHRs data sharing in a health care organization, the two cases are considered such as *simple users case* (i.e., PHR owner has never changed access policy for his users yet) and *policy revocation case* (i.e., PHR owner has changed access policy for his users). In this paper, traditional CP-ABE is used for simple users case and proposed scheme is used for policy revocation case. The PHRs owner may be a data administrator of the whole health care organization who manages PHRs. User may be anyone who is interested in different fields of health care organization (i.e., researchers, staffs, physicians, lab members, nurse, hospital head, and so on.). For this PHRs data sharing, traditional CP-ABE uses a symmetric secret key for encryption/decryption phase. When a policy revocation occurs, proposed scheme uses an updated secret key which is generated by TA according to a new access policy. Both of schemes use an Advanced Encryption Standard (AES) function to encrypt/decrypt PHRs data by using a different secret key. This paper presents a comparison for time measurements of both schemes. Different procedures of encryption, decryption, and key generation algorithms for both schemes are explained in section 4.2. This paper emphasizes on the attributes policy management, key management and supports a flexible policy updating access control. It considers to reduce encryption/decryption times comparing with traditional CP-ABE. Section 2 discusses the related work for ABEs literature reviews. Section 3 describes preliminaries for this paper. Section 4 presents a CP-ABE scheme with policy revocation. Section 5 includes experimental results. Section 6 describes conclusion and further extensions, and finally includes the references.

2. Related Work

Researchers have described the problems occurred in ABE schemes in various ways. Bethencourt et al. proposed CP-ABE by additional consideration for a delegation on an essential attribute structure [1]. They have improved ABE features but they had limitations that it was proved secure under the generic group heuristic and has not considered policy revocation yet [1]. Li et al. studied a survey on ABE scheme of data access control in cloud computing [6]. They listed some unsolved issues of existing schemes such as key management, flexible access and efficient user revocation challenges. They proposed a new scheme Categorical Heuristics on ABE (CHABE). CHABE describes a message and a predicate over the universe of attributes. The attributes set satisfies the predicate, endorsed the message. However, it needs to keep the predicate and message pairs over the universe of attributes in database on server [6]. Yu et al. proposed a combining technique of ABE, proxy re-encryption, and lazy re-encryption technique to achieve a fine-grained data access control in cloud computing [12]. It had multiple system operations and computation on cloud servers which is proportional to the number of system attributes. Ibraimi et al. proposed an encryption scheme for a secure policy updating [5]. They have shown an open problem to provide security proof and to break that scheme for reducing a well-studied complexity-theoretic problem. Wungpornpaiboon and Vasupongayya proposed two-layer ciphertext-policy attribute-based proxy re-encryption for supporting PHR delegation [11]. In [11], the encryption layer is divided into two layers such as inner and outer layer. The inner layer is possessed by data owner and the delegation is processed by satisfying an access structure in the outer layer. Chen and Ma proposed efficient decentralized attribute-based access control for cloud storage with user revocation [2]. It did not need any central authority and coordination among multiple authorities. The authors proposed to consider the user revocation for more practical. Mo and Lin proposed a dynamic re-encrypted ciphertext-policy attributed-based encryption scheme for cloud storage [9]. The authors proposed to consider for re- encryption the ciphertext by using re-key in case of attribute revocation or delegation by delegator. In [9], the re-encryption case was moved to the cloud side to make the data management of the data owner simpler. If that scheme is used, user needs to trust the cloud side. Li et al. proposed flexible and fine-grained attribute-based data storage in cloud computing [7]. In [7], the authors proposed a fine- grained access control (ABE) scheme with efficient user revocation for cloud storage system. The issue of user revocation could be solved by introducing the concept of user group. When any user leaves, the group manager updates users' private keys except for those who have been revoked [7]. In [8], Myint et al. proposed a flexible policy updating access control scheme for cloud storage but it has still an ongoing work for key exchanges and analysis on conventional ABEs. Cui et al. [3] introduced an expressive CP-ABE with partially hidden access structures. Each attribute is divided into an attribute name and an attribute value, and attribute values of the attributes in an access structure are not given in the ciphertext [3].

3. Preliminaries

This section initially describes a number of concepts that provides the basis for proposed scheme.

3.1. CP-ABE Scheme

The four polynomial time algorithms in CP-ABE are as follows:

- Setup (λ, U): This algorithm takes as input the initial information λ such as security parameter and attributes universe description U, and outputs a public key PK and master secret key MK.
- Encrypt (PK, M, $A_{C\text{-}CP}$): This algorithm takes as input PK, plaintext message M and attributes access policy $A_{C\text{-}CP}$. It outputs the ciphertext C associated with $A_{C\text{-}CP}$.
- KeyGen (PK, MK, A_u): This algorithm takes as input PK, MK and access policy of user A_u then outputs a secret key SK.
- Decrypt (PK, SK, C): This algorithm takes as input PK, SK and C then outputs the plaintext message M if and only if A_u satisfies $A_{C\text{-}CP}$ associated with the ciphertext C.

The above algorithms are illustrated in Figure 2. For KP-ABE illustration in Figure 1, $A_{u\text{-}KP}$ is denoted by an access policy of user. A_C is denoted by a descriptive attributes set for a data owner. In KP-ABE, A_C needs to satisfy the structure of $A_{u\text{-}KP}$ as shown in Figure 1. CP-ABE improves the limitation of KP-ABE scheme. In CP-ABE, the data owner has full right on defining access policy before encrypting the message.

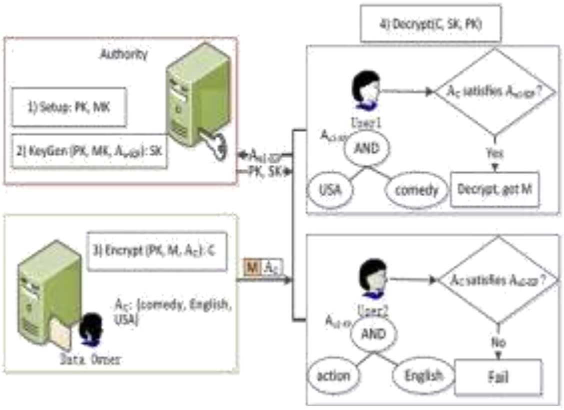


Figure 1. KP-ABE illustration [6]

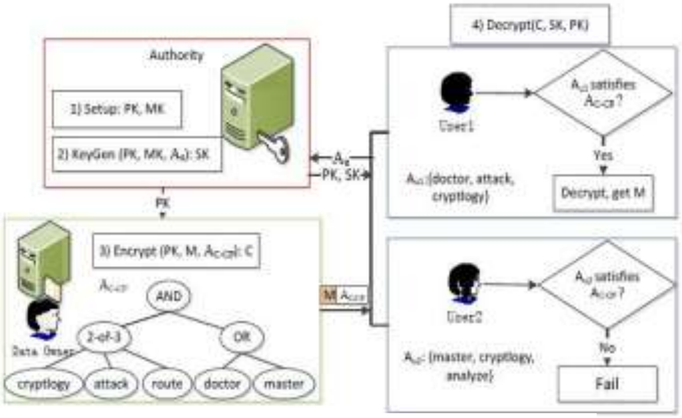

4) Decrypt(C, SK, PK)
Authority
1) Setup: PK, MK
2) KeyGen (PK, MK, A_u): SK
A_u
PK, SK
PK
3) Encrypt (PK, M, A_{C-CP}): C
A_{C-CP}
AND
2-of-3
OR
cryptlogy
attack
route
doctor
master
Data Owner
M A_{C-CP}
User1
A_{u1}:{doctor, attack, cryptlogy}
A_{u1} satisfies A_{C-CP}?
Yes
Decrypt, get M
User2
A_{u2}: {master, cryptlogy, analyze}
A_{u2} satisfies A_{C-CP}?
No
Fail

Figure 2. CP-ABE illustration [6]

3.2. Bilinear Maps

The proposed scheme is based on pairings over groups of prime order. Let G_0 and G_1 be two multiplicative cyclic groups of prime order p, g be a generator of G_0, and Z_p be the additive group associated with integers from {0, … , p-1}. A pairing or bilinear map e: $G_0 \times G_0 \rightarrow G_1$ satisfies the following properties:

1. Bilinearity: for all u, v $\in G_0$ and a, b $\in Z_p$, we have $e(u^a,v^b) = e(u, v)^{ab}$.
2. Non-degeneracy: $e(g, g) \neq 1$. Observe that bilinear map also enjoys the symmetry property, i.e. $e(g^a,g^b) = e(g,g)^{ab} = e(g^b,g^a)$. Group G_0 is said to be a *bilinear group* if the group operation in G_0 and the bilinear map e: $G_0 \times G_0 \rightarrow G_1$ can be computed efficiently.

3.3. Access Tree

Another important concept used in this paper is the concept of an access tree. Let T be an access tree associated with an access policy. A leaf node k in the access tree T represents an attribute from the attribute set $w \in \Omega$, where Ω is a universe of attributes. A non-leaf node k in T represents a threshold gate, which is described by its child nodes and a threshold value. Let num_k be the number of children of a node k and T_k be its threshold value, then $0 < T_k < num_k$. If $T_k = 1$, then k corresponds to an OR gate; if $T_k = num_k$, the node k is an AND gate. For leaf nodes, $T_k = 1$.

4. CP-ABE Scheme with Policy Revocation

This section describes the system structure and system algorithms to implement the proposed methods.

4.1. System Structure

The proposed system structure is shown in Figure 3.

The four entities in proposed system structure are as follows:

- Trusted Authority (TA): An entity which is trusted by all other participating entities in this system. It is responsible for issuing keys to the users upon valid requests.
- Data Owner (DO): The entity who owns data and encrypts those data.
- Data User (DU): The entity who would like to access encrypted data with proper authorization.
- Cloud Storage Provider (CSP): The entity that will provide storage service to store encrypted data.

In Figure 3, PHRs administrator or a DO encrypts each PHR content associated with each policy respectively. The ciphers of PHRs are stored in a cloud server which is a CSP. User or a DU tries to access PHR cipher by proving his credential attributes. A sample PHR training dataset is used in proposed system. A PHR training data consists of an attributes set in which PatientID, Name, NRC, Address, Phone, Disease, Hospital, PolicyID, RevokedPolicyID, and so on. Among these attributes, PolicyID is used for an access control to grant or deny users' accesses. PolicyID is a unique identity number which represents a threshold of three attributes per policy. Each policy consists of the three attributes such as Role, Field, and Hospital. For example, PHRs administrator encrypts a PHR content 'XXX' by defining PolicyID = '15'. Suppose that PolicyID = '15' represents a threshold for " Role = 'Lab member', Field = 'Allergy and Immunology', and Hospital = 'SSC' ". If user has a role of 'Lab member', 'Allergy and Immunology' field, and hospital name 'SSC' in his threshold attributes as in PolicyID = '15', he can decrypt the cipher of 'XXX'. If PHRs administrator changes any attribute in PolicyID = '15' for 'XXX', it means that PolicyID = '15' is updated to another policy identity number for 'XXX'. To prevent collusion attack, the users associated with an old policy '15' for 'XXX' who have to be revoked.

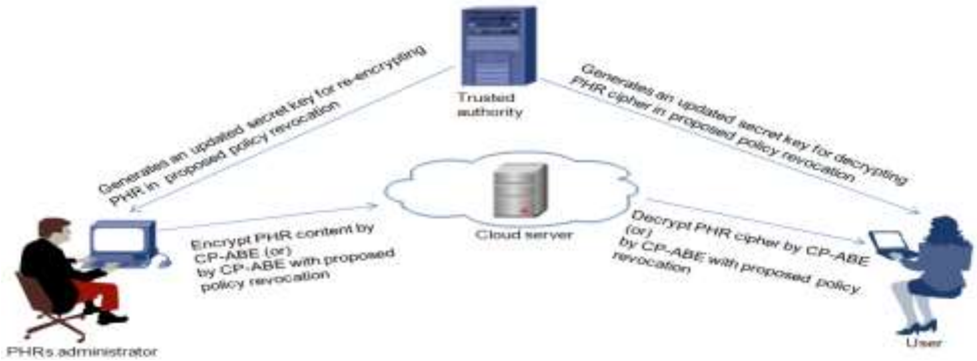

Trusted authority
Generates an updated secret key for re-encrypting PHR in proposed policy revocation
Generates an updated secret key for decrypting PHR cipher in proposed policy revocation
Cloud server
Encrypt PHR content by CP-ABE (or) by CP-ABE with proposed policy revocation
Decrypt PHR cipher by CP-ABE (or) by CP-ABE with proposed policy revocation
PHRs administrator
User

Figure 3. A scenario of PHR data sharing by using traditional CP-ABE and CP-ABE with proposed scheme

4.2. Proposed Algorithms

Algorithms for encryption, key generation and decryption in CP-ABE with proposed policy revocation are as shown in Figure 4, Figure 5 and Figure 6 respectively. Table 1 shows symbols and meanings of proposed algorithms.

Algorithm 1: Encryption algorithm for both simple user case and policy revocation case

```
Input : $M_{PHR}$
Output : $C_{PHR}$
Initialize PHR_Curr_Pol = $Pol_{id}$
If Revo_Pol = NULL && $Status_{id}$ = NULL then
    $C_{PHR}$ = Enc($M_{PHR}$, PHR_Curr_Pol, SK)
Else
    Revo_Pol = PHR_Curr_Pol
    PHR_Curr_Pol = $New\_Pol_{id}$
    $U_{id}$ = $PHR_{id}$
    $Upd_{SK}$ = KeyGen($U_{id}$, $Upd_{level}$)
    $C_{PHR}$ =Enc($M_{PHR}$,Revo_Pol,PHR_Curr_Pol,$Upd_{SK}$)
End If
```

Figure 4. Encryption algorithm

Algorithm 2: Key generation algorithm for updated secret key (policy revocation case)

```
Input : $U_{id}$, $Upd_{level}$
Output : $Upd_{SK}$
Set up a unique value to $Status_{id}$ according to $Upd_{level}$
$SK_{token}$ = $Status_{id}$ + $U_{id}$
$Upd_{SK}$ = GetHashCode($SK_{token}$)
Return $Upd_{SK}$
```

Figure 5. Proposed key generation algorithm by TA

Algorithm 3: Decryption algorithm for both simple user case and policy revocation case

```
Input : $C_{PHR}$
Output : $M_{PHR}$
Initialize User_CurrPol = $U\_Pol_{id}$
If Revo_Pol = NULL && User_CurrPol = PHR_Curr_Pol then
    $M_{PHR}$ = Dec($C_{PHR}$, User_CurrPol, SK)
Else If Revo_Pol = NULL && User_CurrPol ≠ PHR_Curr_Pol then
    Notify "Unauthorized Access!"
Else If Revo_Pol ≠ NULL && User_CurrPol ≠ PHR_Curr_Pol then
    Notify "Unauthorized Access!"
Else If Revo_Pol ≠ NULL && User_CurrPol = PHR_Curr_Pol && $User_{Gid}$ = $Revoked\_User_{Gid}$ then
    Notify "You have been revoked. Don't try a collusion attack!"
Else If Revo_Pol ≠ NULL && User_CurrPol = PHR_Curr_Pol && $User_{Gid}$ ≠ $Revoked\_User_{Gid}$ then
    $Upd_{SK}$ = KeyGen($U_{id}$, $Upd_{level}$)
    $M_{PHR}$ = Dec($C_{PHR}$, Revo_Pol , PHR_Curr_Pol, $Upd_{SK}$)
End If
```

Figure 6. Decryption algorithm

Table 1. Symbols and meanings in proposed algorithms

Symbols	Meanings
PHR	Personal Health Record
M_{PHR}	PHR content data
C_{PHR}	Ciphertext of PHR content
PHR_Curr_Pol	Current policy identity number of M_{PHR}
Pol_{id}	A unique policy identity number for M_{PHR} which is defined by PHR

	owner
Revo_Pol	A policy identity number which is revoked by PHR owner
$Status_{id}$	A unique predefined identity number for updating status according to a policy updating level Upd_{level} (There are four Upd_{level}, so four $Status_{id}$ are predefined for updating status.)
SK	A symmetric secret key (for simple users case)
New_ Pol_{id}	A new unique policy identity number of M_{PHR} (i.e., old policy identity number of M_{PHR} is revoked by PHR owner)
U_{id}	A unique user identity which performs as a temp to keep PHR_{id}
PHR_{id}	A unique identity number of M_{PHR} in PHRs dataset
Upd_{SK}	A unique updated secret key (for policy revocation case)
KeyGen	Key generation function by TA
Upd_{level}	A policy updating level (Four types of Upd_{level} are 'All-Attributes-changes' in policy, 'BelowTheHalf-Attributes-changes' in policy, 'OverTheHalf-Attributes-changes' in policy and 'ByName-changes' in policy.)
Enc	AES encryption function
SK_{token}	A unique secret token key
GetHashCode	A MD5 hash function
TA	Trusted Authority which generates Upd_{SK}
User_CurrPol	Current policy identity number of user
U_Pol_{id}	A unique policy identity number which is proved by user
Dec	AES decryption function
$User_{Gid}$	A unique global identity number of user
Revoked_$User_{Gid}$	$User_{Gid}$ of a revoked user

5. Experimental Results

This section shows experimental results for traditional CP-ABE by comparing with proposed policy revocation in CP-ABE. All of the experimental results are carried out for each PHR data of a sample training PHRs dataset. Training PHR data was explained in previous section. These experiments are configured on a machine of Intel CORE i3 processor, 4GB of RAM, 500GB of HDD and CPU 2.30GHz on Windows7 Ultimate system. It runs on the software version of Microsoft Visual Studio 12.0. Figure 7 shows the measurements of performance evaluation for key generation time, encryption time and decryption time respectively. All of the performance evaluations are measured by an average execution time after testing five times on system algorithms. The key generation time for 3 leaf nodes per policy takes 0.482 seconds by traditional CP-ABE and it takes 0.962 seconds by proposed scheme respectively. The running times for CP-ABE are almost perfectly linear with respect to the numbers of leaf nodes in an access policy. The key generation time for CP-ABE with proposed scheme is longer than traditional CP-ABE because proposed scheme includes an extra consideration to generate an updated secret key for detecting and protecting revoked users according to policy revocation as shown in Figure 7(a). However, longer key generation time is not a weakness for proposed scheme because both encryption and decryption times of proposed scheme are less than traditional scheme. The encryption and decryption times per 10 leaf nodes are 0.37 seconds and 0.34 seconds by traditional CP-ABE. In proposed scheme, the encryption and decryption times per 10 leaf nodes are 0.34 seconds and 0.311 seconds. In proposed algorithms, both encryption and decryption algorithms firstly call the key generation algorithm, secondly take a corresponding key either from TA (in proposed scheme) or from a simple key generator (in traditional scheme). Thirdly, call AES function with the help of CP-ABE. Fourthly, AES inputs that corresponding secret key to do encryption/decryption, and AES transforms it to a system secret key, and then execute corresponding outputs. According to the difference between an old secret key and an updated secret key, the execution times for encryption/decryption are also changed in both schemes. Hence, proposed scheme saves time consuming on overall evaluation for encryption/decryption outputs (which includes calling key generation phase and returning a key) as shown in Figure 7(b) and 7(c). It supports a flexible policy revocation control for user and PHRs owner and it also performs key exchange management, policy management and revoked user detection.

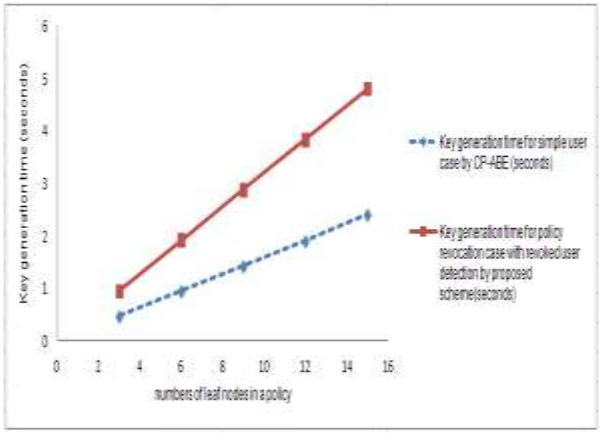


Figure 7(a). Key generation time

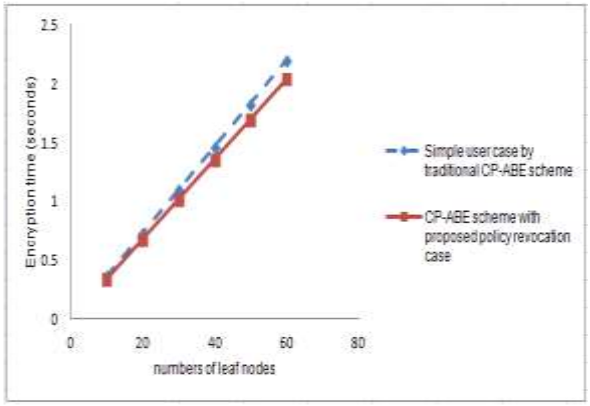


Figure 7(b). Encryption time

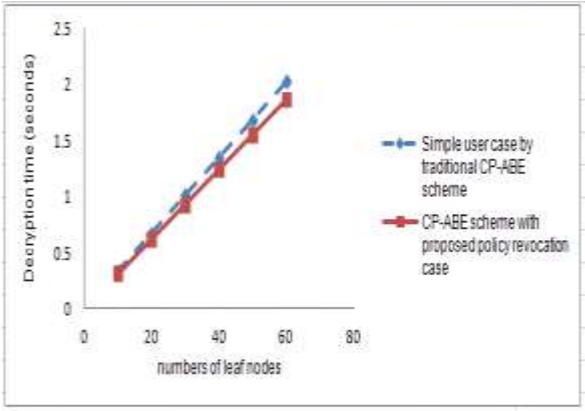


Figure 7(c). Decryption time

6. Conclusion and Future Work

The proposed policy revocation scheme adapts and solves the key management problem for CP-ABE. It focuses on the efficient access policy updating by data owner according to new access policy or policy revocation. It considers generating an updated secret key for encrypting/decrypting the updated PHR. It intends to be more flexible policy management and overall time safe by doing full right on data owner. It is going to study multi authority domains for data protection in cloud storage. As a future work, it is going to do performance analysis by comparing proposed scheme with the existing enhanced CP-ABE schemes.

7. References

[1] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-Policy Attribute-Based Encryption", in Proc. of IEEE Symposium on Security and Privacy (SP'07), 321-334, IEEE Computer Society Washington, DC, USA, May 20-23, 2007.

[2] J. Chen, and H. Ma., "Efficient Decentralized Attribute-based Access Control for Cloud Storage with User Revocation", in Proc. of IEEE International Conference on Communications (ICC), Sydney, NSW, Australia, 3782-3787, Jun. 10-14, 2014.

[3] H. Cui, R. H. Deng, J. Lai, X. Yi, and S. Nepal, "An Efficient and Expressive Ciphertext-policy Attribute-based Encryption Scheme with Partially Hidden Access Structures, revisited", in Proc. of Computer Networks, 133(2018), 157-165, Mar. 14, 2018.

[4] M. George, C. S. Gnanadhas, and S. K, "A Survey on Attribute Based Encryption Scheme in Cloud Computing", in Proc. of International Journal of Advanced Research in Computer and Communication, 2(11), 4408-4412, Nov., 2013.

[5] L. Ibraimi, M. Asim, M. Petkovic, and B. Waters, "An Encryption Scheme For A Secure Policy Updating", in Proc. of the 5th International Conference on Security and Cryptography(SECRYPT 2010), Athens, Greece, Jul. 26-28, 399-408, 2010.

[6] T. Li, L. Hu, Y. Li, J. Chu, H. Li, and H. Han, "The Research and Prospect of Secure Data Access Control in Cloud Storage Environment", in Proc. of Journal of Communications , 10(10), 753-759, Oct., 2015.

[7] J. Li, W. Yao, Y. Zhang, H. Qian, and J. Han, "Flexible and Fine-Grained Attribute-Based Data Storage in Cloud Computing", DOI:10.1109/TSC.2016.2520932, in Proc. of IEEE Transactions on Services Computing, 10(5), 785-796, Sept.-Oct. 1, 2017.

[8] P. W. W. Myint, S. Z. Hlaing, and E. C. Htoon, "An Encryption Access Control Scheme for Flexible Policy Updating in Cloud Storage", in Proc. of the 15th International Conference on Computer Applications(ICCA), Yangon, Myanmar, 28-33, Feb. 16-17, 2017.

[9] L. Q. Mo, and F. Y. Lin, "A dynamic re-encrypted ciphertext-policy attributed-based encryption scheme for cloud storage", in Proc. of IEEE 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Guangdong, China, 14-19, Nov. 8-10, 2014.

[10] C. Vinoth, G. R. A. Raman, "A Survey on Attribute Based Encryption Techniques in Cloud Computing", in Proc. of International Journal of Engineering Sciences & Research Technology, 4(1), 494-497, Jan., 2015.

[11] G. Wungpornpaiboon, and S. Vasupongayya, "Two-layer Ciphertext-Policy Attribute-Based Proxy Re-encryption for Supporting PHR Delegation", DOI: 10.1109/ICSEC.2015.7401447, in Proc. of International Computer Science and Engineering Conference (ICSEC), 23-26, Chiang Mai, Thailand, Nov., 2015.

[12] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing", in Proc. of IEEE Transactions on Parallel and Distributed Systems, 24(1), 131-143, Jan., 2013.

Automatic Adjustment of Consistency Level by Predicting Staleness Rate for Distributed Key-Value Storage System

Thazin Nwe [1], Junya Nakamura [2], Ei Chaw Htoon [1], Tin Tin Yee [1]
[1]*University of Information Technology, Myanmar*
[2] *Toyohashi University of Technology, Japan*
thazin.nwe@uit.edu.mm, junya@imc.tut.ac.jp, eichawhtoon@uit.edu.mm, tintinyee@uit.edu.mm

Abstract

Nowadays, Distributed Key-Value storage is extremely useful in almost every large system. Most of the data management systems use Distributed Key-Value database in order to manage and process large volume of data in real time. In such databases, Apache Cassandra is a peer-to-peer architecture which any user can connect to any node in any data center and can read and write data anywhere. Most of these systems usually select a fixed number of replicas of read/write requests in Key-Value storage. When the more replicas a read request chooses, it may increase the response time and reduce the system performance. Consistency in Key-Value data storage systems requires any read operation to return the most recent written version of the content. The system provides transaction services which provide NoSQL databases with enhanced consistency by varying the read and write quorum size. The proposed approach tends to automatically select the minimum number of consistent replicas by predicting the staleness rate on adjusting the consistency level.

Keywords- Distributed Key-Value Storage, Consistency Level, Quorum, Apache Cassandra, Staleness rate

1. Introduction

Replication is a widely used technology in distributed Key-Value storage systems to achieve data availability, durability, fault tolerance and recovery. In these systems, maintaining data consistency of replication becomes a significant challenge. Although many applications benefit from strong consistency, latency sensitive applications such as shopping carts on e-commerce websites choose eventual consistency [6]. Eventual consistency is a weak consistency that does not guarantee to return the last updated value [8]. Eventually consistent systems are high operation latencies and thus in bad performance. Achieving high throughput and low latency of responses to client requests is a difficult problem for cloud services. To fix these issues, a consistent replica selection process needs to include mechanisms for estimating the latency when processing requests.

Therefore, this system proposes a consistent replica selection approach to read/write access to distributed key-value storage systems by encoding and decoding data onto Distriuted Hash Table (DHT).

This approach can determine the minimal number of replicas for reading request needs to contact in real time by defining the consistency level (one, two, quorum, local quorum, etc.). Depending on these consistency levels, the system can choose the nearest consistent replicas using replica selection algorithms. By using these algorithms, the system will improve the read/write execution time of defining the consistency levels and reduce the read/write latency cost of choosing the nearest consistent replicas.

There are two parts in the system. These are read and write operations for automatic adjustment of the Consistency Level of Distributed Key-value Storage by a Replica Selection Approach. For the write execution time, data is distributed among the nodes in the cluster using Distributed Hash Tables (DHT) that the atomic ring maintenance mechanism over lookup consistency. DHT is mentioned in Section 5. To get the consistent data for the read performance, two algorithms are proposed to Section 4.

2. Related Work

Performance and reliability of quorum based distributed Key-Value storage systems [1, 5] are proposed in the literature. H. Chihoub et al. [3] proposed an estimation model to predict the stale read and the system adjusted replica consistency according to the application requirements. Harmony uses a White box model uses mathematical formula derivation to choose the replicas numbers of each request. To select the number of replicas to be involved in a read operation necessary, this model finds the stale read rate smaller or equal to the defined threshold value. However, since there are so many factors that can impact the result and lots of those factors change in real time, such white box analysis may not get precise result and the system did not consider the performance of read/write operations. Harmony assumes the request access pattern meets Poisson process. However, Harmony has usage limitation because different application access patterns are different.

In most systems, it defines the rate of stale read that can be tolerated, and then try to improve the system performance as much as possible while still not exceed such stale read rate. However ZHU Y et al. [9] takes another mechanism, the longest response time is defined that it can tolerate and try to enhance the consistency level as much as possible within this time. The read / write access is broken into 6 steps: reception, transmission, coordination, execution, compaction and acquisition, and each of which can further break into smaller steps. Then a linear regression is used to predict the execution time and latency of the next request for each step. When a request comes, it maximizes the number of steps this request cover within the tolerated time, thus achieves the maximize consistency. However, the stale read rate of this system is unpredictable.Tlili et al. [7] proposed that a master peer is assigned by the lookup service of the DHT. The master node holds the last update timestamp to avoid reading from a stale replica. However, this system cannot get the precise result of the consistent data and read/write execution time will reduce compared with encoded data on DHT.

3. Proposed Architecture

In our architecture, a client writes data onto the replicas as the write consistency level, i.e., quorum. In the proposed system, the read and write consistency levels (e.g., one, two, quorum, local quorum; etc.) can be defined. For example, if the replication factor of a cluster (N) is defined as five, the sum of read consistency level (R) and write consistency level (W) is greater than a replication factor of a cluster (N). Consistency level describes the behavior seen by the client. Writing and reading at quorum level allows strong consistency.

Data Encoding is a method of data protection in which data is encoded with redundant data pieces and stored across a set of different locations or storage media. The goal of data encoding is to enable data that become corrupted at some point in the disk storage process to be reconstructed by using information about the data that's stored elsewhere in the array. The encoded data of all the transactions are stored in the block.

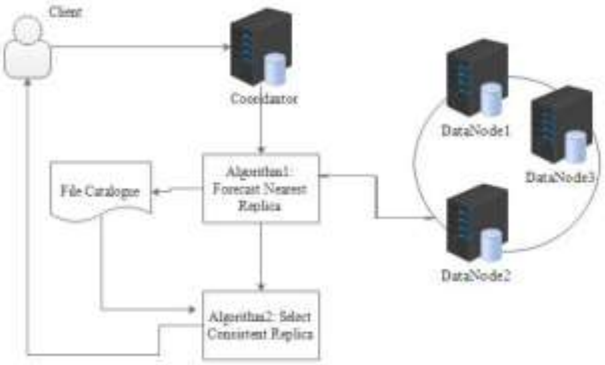

Client
Coordinator
DataNode1
DataNode3
DataNode2
File Catalogue
Algorithm1: Forecast Nearest Replica
Algorithm2: Select Consistent Replica

Figure 1. Consistent Replica Selection Architecture for reading request in Key-Value storage

The system has two parts such as read/write requests. The write requests are incoming to the Coordinator Node. The Coordinator Node performs the encoding operation that divides the data block into m fragments and encode them into n fragments. The fragments created are saved by consistent hashing [10] on different quorum nodes. This means that a coordinator must contact to multiple nodes to decode an original replica. For example, the coordinator has to contact 2 (read consistency level) * 5 (fragments) nodes if RCL=2 and data encoding divides original data into 5 fragments and 2 parities. The acknowledgement of successful write request is sent to the Coordinator Node.

Secondly, when the client reads data, it sends a read request to the Coordinator Node. The Coordinator Node collects the list of DataNodes that it can retrieve data by using the replica selection algorithm described in the next section. When sufficient fragments have been obtained, the Coordinator Node decodes the data and supplies it to the read request. In this case where a client reads from the cluster the file with the read consistency level of five. The Coordinator of the read request retrieves 5*5 (read consistency level *fragments) from data nodes. If two of five nodes fail, the data cannot be lost.

Read and Write operations of the distributed Key-Value storage system to dynamically adjust the consistency level are mentioned in this section. In the write part of the system, Data is distributed among the nodes in the cluster using Consistent Hashing based Function. Consistent Hashing is a widely used technology in distributed key-value storage system. It is a good solution when the number of nodes changes dynamically. And when the virtual node is combined, the load balancing problem will also be solved. Consistent hashing is the algorithm the helps to figure out which node has the key. The algorithm guarantees that a minimal number of keys need to be remapped in case of a cluster sizes change. DHTs characteristically emphasize the following properties:

- Autonomy and decentralization: the nodes collectively form the system without any central coordination.
- Fault tolerance: the system should be reliable (in some sense) even with nodes continuously joining, leaving, and failing.
- Scalability: the system should function efficiently even with thousands or millions of nodes.

In figure 2, when the write requests are incoming to the coordinator node. The coordinator node performs the encoding input data is saved for the Cassandra cluster by consistent hash on different quorum nodes.

When the client reads a file, it sends a read request for the coordinator Node. The Coordinator Node collects the list of DataNodes that it can retrieve data by using the replica selection algorithm described in the next section. When sufficient fragments have been obtained, the Coordinator Node decodes the data and supplies it to the read application request.

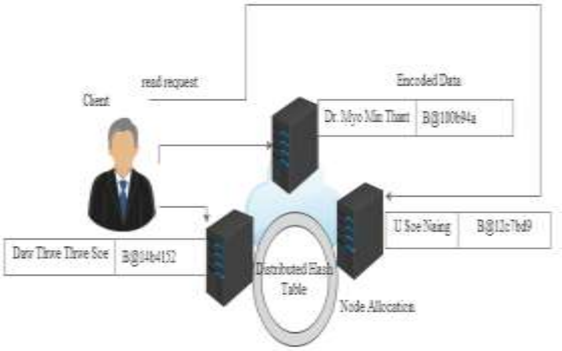


Figure 2. System Architecture

3.1 Storing history Data in Blockchain Architecture

A proof of concept Blockchain system that holds history data containing their timestamps, versions and log data of read/write request time from Key-Value storage cluster.

Blockchain technology promises network-distributed, decentralized and immutable data storage and transaction conduction. The information contained within the blocks make up a database where adding or changing information in that database comes in the form of appending new blocks to the blockchain. A block consists of a header, containing mostly metadata, and transactions that hold the actual blockchain information making up the database.

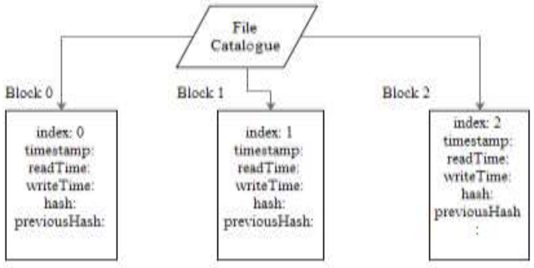


Figure 3. Storage of file catalogue in Blockchain Architecture

4. Algorithm Definition

In algorithm-1, the nodes are mapped to a circle (the value is $0 \sim 2^{32}$) by a specific hash function. There are three nodes (n1, n2, n3), and this algorithm use Consistent Hashing Approach to map three nodes to a circle.

The replica selection algorithm has two parts- (i) searching the nearest replica and (ii) selecting consistent replicas. In algorithm-2, the coordinator node sends the request message to each node who has one or more replicas and latencies of different replicas are listed in the read latency map. And it chooses the lowest latency of replicas of this map.

The Coordinator node executes the replica selection algorithm (algorithm 3) and chooses the consistent replica of the nearest replicas.

```
Input: The nodes of Cassandra Cluster
Output: Distributed Hash Table Nodes
For each n in Cassandra Cluster
  Begin
      For r=1,2,…n do
      Process the hash  function for each node in DHT
        Return DHT nodes
        End for
  End
```

Algorithm 1: Distributed Hash Table Nodes

```
Set lowestLC [] =null; //Initialize return lowest latency replica
For each r in DHT
  Begin
      Set latencyCost=getLatencyCost ($RF_r$, job);
```

```
        If
          (latencyCost<=MAX_VALUE)
        Then
          MAX_VALUE=latencyCost;
  //MAX_VALUE =threshold values
          LowestLC. add ($RF_r$);
          Return lowest LC //nearest replica NR
    End
  End for
```

Algorithm 2: Search the nearest Replica

Search the nearest Replica part is divided into two stages. First, all replicas are sorted based on their physical location, so that all replicas of the same rack and then the same data center as the source are at the top of the list. Second, the latencies are computed from the local node (originator of the query) to all other nodes. If the latency cost is greater than a threshold of the closest node, then all replicas are sorted based on their latency costs. Finally, the top replicas of the list are chosen.

Firstly, total numbers of replicas are listed as input. The threshold value is set at the latency cost. The coordinator node contacts every other replica with request messages. The round tripped to the time it takes from the request until the reply is passed through the following equation (1).

$$T_{total} = \frac{RTT\ request}{2} * Tprocessing + \frac{RTT\ reply}{2} \qquad \textbf{(1)}$$

T_{total} is used to get the latency cost of computing, data nodes in algorithm2. These latency costs are used when the local node needs to forward the client request to other replicas.

And then total times taken from different replicas are listed in latencyCost. Finally algorithm2 returns the list of lowest latency cost of the replicas as output to client.

```
 Input: Nearest Replica NR= {$NR_1$, $NR_2$,… $NR_n$}
 Output: Consistent Replicas
 For each Nearest Replica $NR_i$
   Begin
       Set RCL=2//ConsistencyLevel.QUORUM
       Set noOfConsistentRead = 0
     While (noOfConsistentRead <= RCL)
       If (stalerate <= maxStaleRate)  Then
         consistentRead.add ($NR_i$)
         noOfConsistentRead++;
       Return consistentRead;
   End
End for
```

Algorithm 3: Consistent Replica Selection

In algorithm-3, the set of the nearest replicas is collected as input that comes from the output of algorithm 1 by computing latency costs. And then algorithm 2 sets the read consistency level (RCL) that the client will need the most up-to-date information. Read/Write latencies of different replicas are listed in history file on the coordinator node.

This algorithm determines the number of consistent replica nodes, one read request should select in real-time, according to calculated arrival times of nearest update request and the processing order of the read request and write request in different replicas.

For computing stale rate of algorithm 3, In this section, we will use PBS to simulate more cases of replica number n, read consistency level r, and write consistency level w to prove the conclusion. Bailis et al proposed PBS to predict the consistency [2]. They believed the quorum size impacts the consistency significantly and given a formula to calculate the probability of k-staleness [2]. In staleness estimation, a client estimates the staleness of each node in the quorum, based on the last write date values provided in server, and select only those secondaries whose staleness is less than or equal to maximum staleness rate. The calculation of staleness is the following equation (2).

$$p_{st} = \frac{\binom{N-W}{N}}{\binom{N}{R}} + \sum_{c\in(W,N]} \frac{\binom{N-c}{N}}{\binom{N}{R}} \cdot [P_w(c+1,t) - P_w(c,t)] \qquad \textbf{(2)}$$

For quorum system, consider N replicas with randomly chosen read and write quorums of sizes R and W. The system calculates the probability that the read quorum does not contain the last written version. This probability is the number of the quorums of size R composed of nodes that were not written to in the write quorum divided by the number of possible read quorums.

4.1 PBS <k, t > staleness

PBS <k, t > staleness combines both versions and real-time staleness metrics to determine the probability that a read will return a value no older than k versions stale if the last write committed at least t seconds ago:

A quorum system obeys PBS <k, t> staleness consistency, if, with probability 1 − pskt, at least one value in any read

quorum will be within k versions of the latest committed version when the read begins, provided the read begins t units of time after the previous k versions commit.

The probability of inconsistency is calculated by the following equation (3):

$$P_{skt} = \frac{\binom{N-W}{R}}{\binom{N}{R}} + \sum_{c \in (W,N)} \frac{\binom{N-c}{R}}{\binom{N}{R}} . [P_w(c+1,t) - Pw(c,t)])k \qquad (3)$$

The staleness result of this equation is calculated in section 5.

5. Experiment and performance evaluation

We evaluate the proposed algorithms and read/write execution time of consistency level by using a Cassandra cluster of VMware Ubuntu 14.04 LTS i386. The processor is Intel (R) Core (TM) i7-4770 CPU @ 3.40 GHz. Installed memory (RAM) is 4.00GB.

Table 1. Hardware Specification and Virtual Environment

Operating System	VMware Ubuntu 14.04 LTS i386
RAM	4.00GB
Hard-disk	195GB
Processor	Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz
Cassandra	version: 1.0.6

The Exchange rate currency is used in Cassandra cluster. Exchange rate currency information is described by Unicode in “currency.csv”. When importing data onto the csv to Cassandra, Java hector code truncates the input csv data with a comma (",") lined by line. And then the output csv data are exported on Cassandra.

Ubuntu14.04 LTS is installed on three servers and one client of the Cassandra clusters. There are 203 rows and 40 columns from csv file are inserted into Cassandra clusters with Consistent Hashing DHT (Distributed Hash Table).

DHT is one of the fundamental algorithms used in distributed scalable systems. DHT deals with the problem of locating data that are distributed among a collection of machines. In the general case, a lookup service may involve full-content searching or a directory-services or structured database query approach of finding data record that matches multiple attributes. Lookup service similar to a hash table: (Key, Value) pairs are stored in a DHT, and any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from keys to values is distributed among the nodes [4].

The encoded data by writing execution time for distributed DHT nodes are better than simple DHT [10]. DHT is a widely-used solution to search the nearest neighbor node that transforms the data to obtain a short code consisting of a sequence bits. Each experiment is repeated five times and the average result is considered for performance evaluation. The staleness rate for consistency label is mentioned in figure 4, 5. If client reads with QUORUM, then 2 nodes will get fresh data, and when the system reads with QUORUM, then the system calculates the staleness rate and choose the replica with minimum staleness rate to get fresh data. The following results are calculated by Probability of Bounded Staleness based on time and version is mentioned in section 4.1.

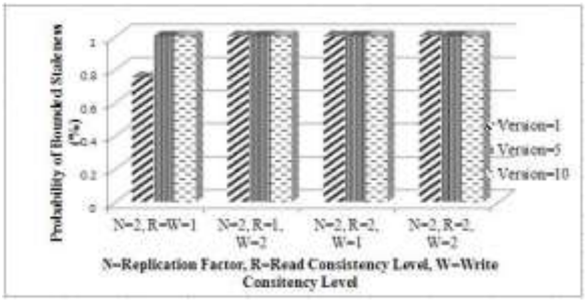


Figure 4. Probability of Staleness rate for Consistency Level on Replication Factor=2

When N=3, R=W=1, this means that the probability of a staleness rate within 1 version is 0.75, within 3 versions, 3.38 ends, 5versions, 7.59. When N=3, R=2, W=2, these probabilities decreases that within 2 versions is 1.44, within 3 versions is 1.75, and within 5 versions is 2.49.

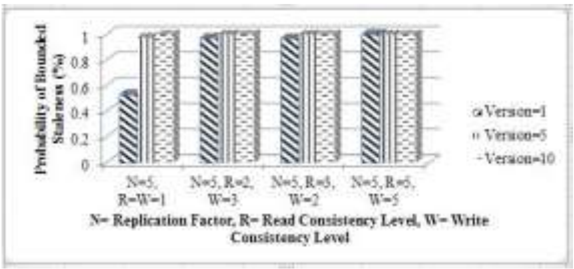


Figure 5. Probability of Staleness rate for Consistency Level on Replication Factor=5

When N=5, R=3, W=5, this means that the probability of a staleness rate within 2 versions is 2.37, within 3 versions, 3.64 and, 5 versions, 8.60. When N=5, R=3, W=8, these probability increases that within 2 versions is 1.71, within 3 versions is 2.24, and within 5 versions is 3.83. The system defines the decision how to define the read/write consistency level by predicting the staleness rate.

For QUORUM of both read & write requests, the system gets strong consistency per following equation

R + W > N, where R - read replica count, W - write replica count, and N - replication factor.

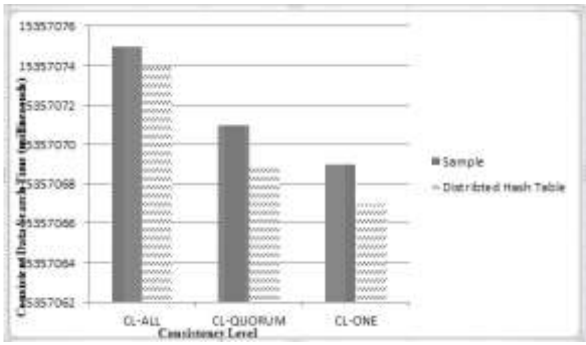


Figure 6. Search Time for consistent data in DHT

In figure 6, the system compares the search time of consistent data onto DHT and sample search. By using DHT, the closest node to the client is looking for the data with the identifier. If the data doesn't exist, this node which sends it to other nodes which send it back to sender node. Therefore, the system reduces the size of the jump at each step and are faster than sample DHT.

6. Future Work

In the future, the proposed algorithms will be validated with more datasets in distributed Key-Vale storage system and the performance of the system is measured by storing log data in Blockchain Architecture.

7. Conclusion

The proposed system presents the prediction of the staleness rate for Key-Value data storage of financial data onto DHT by adjusting the consistency level. In defining these consistency levels, two algorithms are proposed to choosing the consistent replicas of different clusters by searching the nearest replica and selecting the consistent replica. The proposed algorithms can determine the minimal number of consistent replicas of reading request needs to contact in real time and thus improve the system performance as a result of reduced read/write execution time. The system examined a real-time data replicated storage system in which content updates are replicated and stored in a quorum systems. Either a write or read request goes to all the nodes in the system, and it is considered complete once there are at least w or r responses. Assuming that the write delay dominates the latency, the freshness of the replicated storage system is measured by the average age of the content returned by a read at any time t.

8. References

[1] Agrawal, D., El Abbadi, A.: The generalized tree quorum protocol: An efficient approach for managing replicated data. ACM Trans. Database Syst. 17(4), 689.

[2] Bailis P, Venkataraman S, Franklin MJ, Hellerstein JM, Stoica I. Probabilistically bounded staleness for practical partial quorums. Proc VLDB Endowment. 2012;5(8):776–787.

[3] H. Chihoub, S. Ibrahim, G. Antoniu and M. S. Perez, "Harmony: Towards Automated Self-Adaptive Consistency in Cloud Storage", IEEE International Conference on Cluster Computing, September 24-28; Beijing, China , 2012.

[4] Jingdong Wang, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen, "A Survey on Learning to Hash", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 13, NO. 9, APRIL 2017.

[5] Malkhi, D., Reiter, M.: Byzantine quorum systems. pp. 569_578. TOC'97,ACM(1997),http://doi.acm.org/10.1145/258533.258650.

[6] P. Garefalakis , P. Papadopoulos, I. Manousakis, and K.Magoutis, "Strengthening Consistency in the Cassandra Distributed Key-Value Store", International Federation for Information Processing 2013.

[7] Tlili, M., Akbarinia, R., Pacitti, E., Valduriez, P.:Scalllable P2P reconciliation infrastructure for collaborative text editing, Second International Conference on Advances in Database Knowledge and Data Applications (DBKDA), pp. 155_164, April 2010.

[8] W. Vogels, "Eventually consistent", CACM, 52:40–44, 2009.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[9] Y. Zhu and J. Wang. Malleable, "Flow for Time-Bounded Replica Consistency Control", OSDI Poster, October 8-10; Hollywood, USA, 2012.

[10] Thazin Nwe, Tin Tin Yee, Ei Chaw Htoon, Automatic Adjustment of Read Consistency Level of Distributed Key-Value Storage by a Replica Selection Approach, ITC-CSCC Bangkok, vol.33, pp. 740-741, 2018.

Data Mining

Towards the Discovery of Genuine Social Groups from Mobility Data

Htoo Htet Aung
edward@aidatech.io
AIDA Technologies

Nay Aung Lwin
lwin.nayaung@gmail.com
EIE Co. Ltd

Phyo Phyo San
phyophyo.san@suncorp.com.au
Suncorp Group

Abstract

Extraction of social groups from human mobility datasets has been regarded as convoy mining problem since a convoy (defined as a group of people which are spatially close to each other across time) is customarily assumed to represent a social group. However, social groups cannot be modelled trivially as a convoy as empirical evidence suggests convoy mining will report many 'false positives' and miss some 'false negatives'. We propose a two-step method to discover social groups from human mobility data in real-time. To the best of our knowledge, this paper is the first attempt to discover genuine social groups from mobility data. Experiments on the real-life dataset indicate that our two-step approach can accurately and efficiently discover genuine social groups.

Keywords- Real-time Data Mining, Mobility Data

1. Introduction

Human mobility data collected by systems like [1] can be analyzed either off-line or in real-time to obtain actionable insights, useful information and knowledge. For instance, Aung and Tan [2] has extracted frequently used routes from GPS traces of trucks and human users.

An important mobility analytics on pedestrian data is to extract social connection among the people tracked in the dataset as this information can lead to better operations of venues. For example – venue operators can notify a social group (a family) that someone from their group (an underage child) has gone missing soonest possible as a real-time social-group tracking system can monitor social-groups in public places.

Another useful application of social group information is in understanding of the population under analysis. For example, operators of museums and shopping malls will have a better understanding of the demographics of their visitors (families or singles etc.), which is essential to perform targeted advertisement or promotions. It also plays an important role in crowd control as noted in [3].

A naive approach to extract social groups from human mobility data is to model a social group as a convoy — traditionally defined as a group of users moving together — and employ one of the convoy mining techniques [4, 5, 6] to obtain social groups. However, consider the three independent pedestrians ('*a*', '*b*' and '*c*') walking down a narrow corridor in Fig. 1. Since they are moving together, the naive approach will detect them as a convoy and wrongfully report them as a social group even though they are not. We will term such instances 'false positive.' On the other hand, the family ('*p*', '*q*' and '*r*') is a social group yet the traditional convoy mining algorithms will not capture them as a convoy as the child '*p*' does not always move together with her parents, '*q*' and '*r*'. The naive approach will, therefore, will not report this family — a 'false negative.' The naive approach can capture only the couples ('*x*' and '*y*') as they move together.

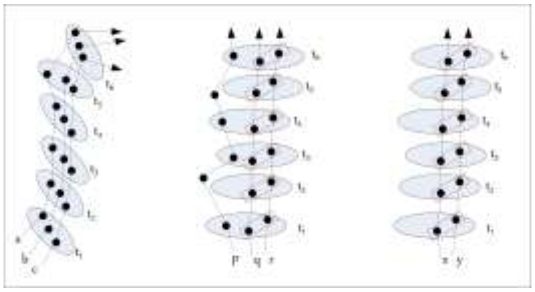


Figure. 1. An Example of Mobility Dataset

The false-positives and the false-negatives negatively impact the application in question. Take, for example, the three independent pedestrians ('*a*', '*b*' and '*c*') in Fig. 1. Alerting the group that their members are lost after each pedestrian take separate ways will be a waste. Likewise, only detecting the parents ('*q*' and '*r*') as a social group and not alerting when the child ('*p*') wonders away from the group for an extended period is not desirable.

To capture these 'false negatives' as social groups, Aung and Tan [7] introduced the notion of 'dynamic convoys' to capture the family in Fig. 1 correctly as a social group yet their approach still unable to weed out the false positives such as the group of '*a*', '*b*' and '*c*'. Zanlungo *et al.* [3] suggested that a genuine social group may be distinguishable from group of independent people moving together by using group features such as group formation and group velocity. For instance, from the fact that '*a*', '*b*' and '*c*' does not have a group formation (they don't have a member keeping in view of other members) and their high velocities, it is subtly hinted that they are not very likely to be a social group. In contrast, the couple and the parents in the family maintains a group formation ('*q*' and '*x*' maintain '*r*' and '*y*' respectively in their field of view by falling behind slightly as highlighted in rectangles). However, their findings are limited to simulating crowd dynamics and are not readily adaptable as models and algorithms for discovering social groups.

Indeed, users in a social group exhibit far more complex movement behaviors than a simple model like a convoy can characterize. Empirical evidence (See Sect. 6 for details) that naively modelling of social groups as convoys does not yield satisfactory results in accuracy. In addition, we postulate that a social group may exhibit different movement patterns in different environments, i.e. a group formation in a crowded corridor may be different from that of the same group in a park. Hence, a single rule-set or a model to describe movement patterns of social groups is impractical. Therefore, we propose to employ machine learning techniques to model the genuine social group by learning the movement behaviors of known social groups, after which the model can be used to effectively discover social groups in a similar dataset.

In this paper, we:

1) **report that a traditional convoy do not necessarily indicates a social group**. In other words, discovery of social groups using convoy mining methods will contain ‘false positives’ and miss ‘false negatives’.

2) **proposed a two-step framework to extract genuine social group from human mobility data**. Our two-step framework consists of a mining step to get potential social groups and a classification step to determine which potential social group is a social group.

To the best of our knowledge, this is the first work that addresses the discovery of genuine social groups both accurately and efficiently. The first step in our two-step framework is based on the dynamic convoy mining algorithm [7] and can capture groups which do not always move together. The second step is a machine learning based classifier that identify social groups from the groups mined in the first step. It currently uses gradient boosting machine classifier we built using multiple weak classifiers. Experiment results show that our two-step framework is more efficient and accurate than the convoy approaches.

2. Related Works

In this section, we will discuss variants of convoy models and background on classification algorithms.

Flocks. Gudmundsson and Kreveld [4] defined a Flock pattern *f lock* (*m*, *r*, *w*) as a pattern formed by a set *G* of at least *m* objects staying in a *moving* circle of radius *r* for at least *w* consecutive time-stamps. They reported that the complexity to compute the all of flock patterns is NP-Hard. Vieira *et al*. [5] reported polynomial-time algorithms to find Flock instances of fixed durations.

Convoys. Flock model has a short-coming —a circular area with maximum radius *r* cannot fit more than a certain number of members — termed as lossy-flock problem by Jeung *et al.* [6]. Using the definition of density-connectivity (two density-connected objects have a chain of dense neighbors that connect them) from [9], they defined a Convoy as a group of at least *m* objects being density-connected with each other throughout *w* consecutive time-points, where *m*, *w*, and the two DBSCAN parameters ε and *min-pts* = *m* are provided by the user. Since a convoy can occupy an arbitrary shaped region in its lifetime, there is no lossy-flock problem.

Dynamic Convoys. Aung and Tan [7] proposed Dynamic Convoy to mitigate shortcomings of convoy models requiring to have all its members spatially close from convoy’s formation to its disintegration. For given parameters: *m*, *k* and *w*, a set of moving objects *D* forms a dynamic convoy during a period $P = [t_{\text{start}} (D), t_{\text{end}} (D)]$ if it contains (i) at least *m* persistent-members, all of which are density-connected in each time-stamp *t* in *P* and (ii) zero or more dynamic members, each of which must be density- connected with the persistent-members at least *k* times for any *w* sliding window in *P*. The definition of Dynamic Convoy allows dynamic members to move away from the main body and, thus, reduces “false-negatives”.

Background on Classification. Traditional tools and notions used in query processing become impractical and inadequate when the underlying semantics are complex to capture in a simple rule/query-model. This is more apparent in social group discovery as the same social group may behave and move in different ways in different applications’ environments. Machine Learning techniques can overcome this issue by building the model from data.

Supervised machine learning is a type of machine learning algorithm that learns from the training dataset, to classifies unseen instances in the test dataset. It works in two phases – training phase, where the machine learns from the training dataset to produce a model and the prediction phase, where the machine uses the model to classify the test data. An instances of supervised machine learning technique is the decision tress based gradient boosted machine (GBM).

GBM Model. McCaffrey *et al.* [9] proposed GBM model, which is a piecewise constant model for prediction of dichotomous outcomes. Initially, it starts with a single simple regression tree while other trees are constructed and added at each new iteration, in which the new tress is determined to provide the best fit to the residuals of the model from the previous iteration and provides the greatest increase to the log likelihood for the data. Exploiting the connection between boosting and optimization, Friedman [10] introduced the gradient boosted machines. GBM can build a strong classifier by combining weak classifiers.

3. Problem Definition

Mobility Dataset — For a given period of time $T = \{t_1, t_2, ..., t_\tau\}$ and a set of users $U = \{u_1, u_2, ..., u_n\}$, a set $\mathbb{D}$ of records of the form $\langle u,t,x,y \rangle$, where $u \in U$, $t \in T$ and $(x,y) \in \mathbb{R}^2$, is a Mobility Dataset. In a Mobility Dataset, each record $\langle u,t,x,y \rangle$ represents user u is at location (x,y) as sampled on time-stamp t. Without loss of generality, we assume that we can access the mobility dataset in ascending order of time-stamps. This reflect real-world application scenarios with streaming data. Fig. 1 and Fig. 2 depict examples of human mobility datasets chronicling movement records of $U_{tr} = \{a,b,c,p,q,r,x,y\}$ and $U_{ts} = \{d,e,f,i,j,s,t,u\}$ respectively. The ovals indicate proximity.

Social Group Information — Given a mobility dataset $\mathbb{D}$, its social group information $\mathbb{G}$ is a set $\{G : G \subset U$ and G is a social group $\}$. For example, consider the mobility dataset in Fig. 2, its social group information contains the family and the couple, $G_{ts} = \{\{s,t,u\},\{i,j\}\}$. Since social associations do not necessarily translate into a set of identifiable movement patterns and vice versa, finding social groups usually is labor-intensive and requires both mobility datasets and other sources of data.

Thus, the discovery of social group – to find all social group information from a given mobility dataset – has become a challenge as the mobility datasets explode.

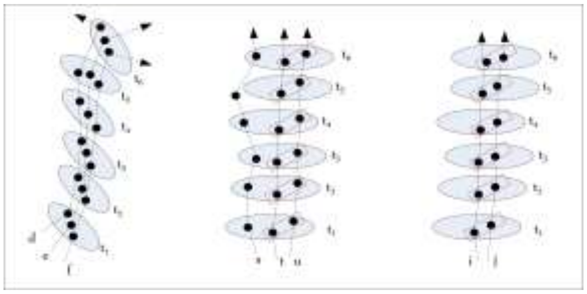


Figure. 2. Mobility Dataset for Test

4. Attempts to Find Social Groups

Convoy — For given parameters: m, k and $w = w_c$ ($m > 1$, $1 \leq k \leq w$), a set of moving objects C forms a convoy in time period P of length w_c if it:

- Contains at least m persistent-members, all of which are density-connected in all t in P and
- Contains dynamic members, each of which must be density-connected with the persistent-members at least k times in P

This definition of convoy is that of a dynamic convoy defined in [8] except that it lasts exactly a period of length w_c. Consider the family $\{p, q, r\}$ in Fig. 1. For $m = 2$, $w = 6$ and $k = 4$, it forms a dynamic convoy as q and r form the main body of the convoy and p is registered as a dynamic member. Using this definition, we can build a program to mine all convoys in a given mobility dataset.

Following the assumption that a social group will be always together, we first attempt to approximate a social group with convoys by regarding each convoy is a social group. Following our running example in Fig. 2, there are three convoys formed by $\{d, e, f\}$, $\{s, t, u\}$ and $\{i, j\}$. Therefore, this approximation method reports three social groups including the group of three independent pedestrians $\{d, e, f\}$. We term such an instance of wrongfully reporting a group with no social connection a 'false positive.' Likewise, since not all social groups move like a convoy, this attempt can also miss to report a genuine social group, which we term as a 'false negative.'

Limitations. In our experiments (see Sect. 6.2), where we compare a human-labelled social groups with convoy results, we discovered that this approximation method yields very few false negatives but many false positives.

Since we observed that the convoy model derived from fixed movement patterns cannot represent a social group, we consider modelling social groups using a machine learning technique. To do this, we compute group features and manually label each group. From the computed features and labels (the training data), machine learning algorithm learns and produces a model.

If we use the mobility dataset shown in Fig. 1 as training data, we need to compute features and mark label for all $G \subset \{a, b, c, p, q, r, x, y\}$. Table 1 shows a few groups along with three features and their labels. The three features are x_1 = whether the group is elongated towards the group direction, x_2 = whether the members keep other members in view. The features computed for the family $\{p,q,r\}$ is $x_1 = N$ and $x_2 = Y$ (in the second row).

Table 1. Examples of Features and Labels

G Group	x_1 (elongated)	x_2 (keep in view)	y
$\{a, b, c\}$	Y	N	N
$\{p, q, r\}$	N	Y	Y
$\{x, y\}$	N	Y	Y

Learning from the training data, the machine will produce a model, using which we can identify social groups from unseen/future test data. In this example, the output model will be "Social groups are groups with $x_2 = Y$ and $x_1 = N$". Applying this rule-set (model) to identify social groups in the test dataset shown in Fig. 2 results in $\{s,t,u\}$ and $\{i, j\}$ correctly outputted as social groups..

Limitations. Although machine learning improves accuracy, it is not practical because we need to calculate features for all groups in the given dataset. The number of all the possible groups are exponential to the number of users. For the test dataset (Fig. 2), we need to calculate features for 247 groups (2^8 = 256 subsets of 8 users minus 9 subsets with size less than 2). Since real-life applications have thousands of users, this method is not feasible.

5. Proposed Framework

Since approximating social groups with convoys is efficient and identifying social groups using models built by machine learning is accurate, we combine the merit of these two methods into a framework to discover social groups. Since convoy approximation usually contain fewer false negatives than false positives, we use it as a filtering step to reduce the number of groups for feature computation to a manageable size. Then, in the next step, we use a machine-learned model to identify social groups.

Box 1. outlines our proposed framework. In training phase, convoys are mined in training dataset D_{tr} (line 1).. Then, features extracted from convoys and labelled groups are used to build a classifier *clf* using GBM (line 2). In test phase, for each sliding time-window, two steps are performed to get the social groups. The first step is mining potential social groups using a convoy mining algorithm (line 4). The second step is calculating features for the potential groups (line 5) and classifying if each potential social group is a genuine social group using the classifier *clf* (line 6). The test phase is designed to work incrementally using a sliding time-window and, hence, our framework can work on a streaming dataset.

```
Input:   Training data, $D_{tr}$ and $G_{tr}$, Convoy parameters, ($e$, $m$, $k$, $w$)
         and GBM learning rate delta and Testing data $D_{ts}$
Output: a set of social groups
  1.  Convoys in Training Data $C_{tr}$ ← Convoy ($D_{tr}$, $e$, $m$, $k$, $w$)
  2.  Features $X$ ← Features ($C_{tr}$), clf ←GBM (delta, $X$, $G_{tr}$)
  3.  for each length $w$ sliding time-window in $D_{ts}$ do
  4.        Potential social groups $G'$ ← Convoy ($D_{ts}$, $e$, $m$, $k$, $w$)
  5.        Features $X'$ ← Features ($G'$))
  6.        Output social groups clf ($X'$)
```

Box 1. Framework to Discover Social Groups

We are going to illustrate how our framework works using the datasets in Fig. 1 and Fig. 2 as training and test datasets. In the training phase, convoys formed by {*a*,*b*,*c*},{*p*,*q*,*r*} and {*x*,*y*} will be mined and features will be calculated for them. Based on the features and labelled data G_{tr} = {{*p*,*q*,*r*},{*x*,*y*}}, the classifier *clf* is built. Once the classifier is trained, the test phase begins. In the first step of test phase, potential social groups G' is mined to find convoys formed by {*d*,*e*,*f*}, {*s*,*t*,*u*} and {*i*,*j*}. In the second step, features of these three groups are calculated. Notice that in contrast to applying a machine learning classifier directly, where features of 247 groups needed to be calculated, our framework only calculates features of 3 groups. These features will be used to identify {*s*,*t*,*u*} and {*i*, *j*} as social groups while {*d*,*e*,*f*} will not be reported.

Mining Potential Social Groups (MPSG). The first step in our proposed framework is to mine convoys to reduce the number of groups, for which feature computation to be performed. We adapt the S^3 proposed in [7] to discover all the convoys as the S^3 algorithm is capable to mine convoys in incremental manner. An alternative candidate is the online flock mining algorithm proposed in [6] but this algorithm has lossy-flock problem and, thus, will introduce false-negatives.

S^3 algorithm requires four parameters, ε, m, k and w_c to control the model or the query of the convoy mining process. All these parameters are intuitive and easy to set.

Identifying Social Groups (ISG). The second step in our proposed framework is to identify genuine social groups from the potential social groups. We choose Gradient Boosting Machine (GBM) for its simple regularization strategy and its ability to build good classifiers from weak classifiers. The regularization of a GBM learning process can be tuned through the shrinkage parameter. Other parameters include number of estimators and minimum sample split to split an internal node.

Features Selection. We calculated 57 features in total reflecting group movement (such as mean group velocity) and relative position/movement angle of group members. Most relative features are based on the movement angle and frame of the group, which are defined as follows.

Let $\mathcal{L}$ = {$\alpha_{1,1}$, $\alpha_{1,2}$, ..., $\alpha_{1,m}$, ..., $\alpha_{n,1}$, $\alpha_{n,2}$, ..., $\alpha_{n,m}$ } represent the movement angles of a group of n users over m time-stamps. The movement angle of a group is *mean_ma* = *arctan*(*mean*(*sin*(α)),*mean*(*cos*(α))).

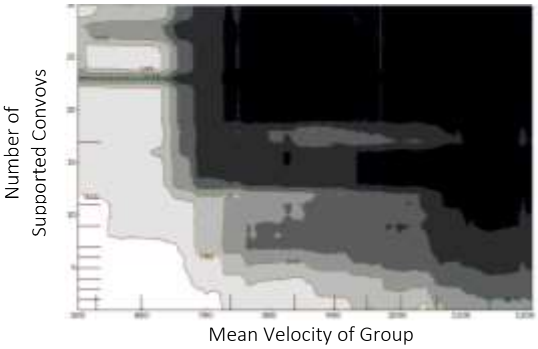
Let θ be the movement angle of a group of users G for a period P, the group-frame of G at time-stamp $t \in P$ is the bounding rectangle covering all the locations of the group members at time t and having sides parallel and orthogonal to the x' axis of the coordinate system rotated by θ from the default (x,y) coordinate system.

From movement angle and group frames of a user group, we compute features. Table 2 tabulates the three most important features we obtained from the model.

Fig. 3 shows the likelihood of a group being a social group given the feature values (darker means higher likelihood). We observe that faster movement and more occurrence of convoys formed by the group also indicate high chance of being a social group. Likewise, groups elongated in the direction orthogonal to the group movement angle is more likely to be a social group Since these observations agree with our observation in the dataset, we deduce that the features we select can differentiate a genuine social group from a mere convoy.

Table 2. Three Most Important Features

No.	Feature	Description
1	Mean_Velocity	Mean velocity of the group
2	Num_Convoys	Number of convoys supporting this group
3	Group_Length	Length of the group-frame parallel to group's mean velocity



Figue 3. Partial Dependence Plot for Features, Mean_Velocity and Num_Supported_Convoys

6. Experiments

We conducted experiments to assess the performance (run-time and accuracy) of our proposed framework against other methods based on Flock and Convoy models. Since our framework can discover social groups in real-time, we tried to compare it against Flock and Convoy models specifically because they have incremental algorithms to mine them in real-time. Flock represents all traditional convoy models while Convoy represents the dynamic convoy model (defined in Sect 4).

We used five sets of human mobility data [11] collected in five different days. Each day of datasets consists of four hours (10:00-11:00, 12:00-13:00, 15:00-16:00, 19:00-20:00) of movement data along with social groups labelled by a human coder [3]. We used two datasets (0424 and 0508) for performance comparison while the rest were used to train the GBM model for our framework. Details of the datasets are given in Tab. 3.

We define the step size of the sliding window for all three methods as 5 second. We chose parameters for Flock and Convoy methods to obtain the best accuracy. For convoy, we set the parameters m = 2, w_c = 6 and k = 5 while for flock, we set similarly as m = 2 and w_c = 6 — a social group should stay together for 6 time-stamps, which are 5 second apart, i.e. 30 second. Clustering parameter for convoys is given as e =1.5m while the disc size for flocks is r = 0.75m. The difference in clustering/disc-size parameters is fair as they reported similar results.

Table 3. A Summary of the Datasets

Name	Num. Objects	Covers	Num. Records	Num. Social Groups
0109	3,774	09/01/13	210, 079	317
0217	7,465	17/02/13	449, 201	1, 221
0324	7,472	24/03/13	449, 804	1, 242
0424	2,750	24/04/13	140, 867	326
0508	2,858	08/05/13	156, 913	384

For the convoy sub-routine in MPSG step in our proposed framework, we chose to set w_c = 3 and k = 2 respectively. This relaxes the convoy selection criteria and introduces a very large number of false positives while reducing a few false negatives. Although this trade-off reduces accuracy of the MPSG step drastically, it increases the overall accuracy of our framework because MPSG step will report less false negatives while the ISG step will discard the false positives MPSG step introduces.

We performed 5-fold cross validation in a small parameter space to find satisfactory parameters for the GBM learner in the initialization phase of our framework. We found that learning-rate g = 0.05 and number of estimators = 150 gives good accuracy performance.

In measuring accuracy of each method, we counted a group reported by a method as a true positive if and only if there is an exact match with a social group in the ground-truth data, i.e. if {p, q, r} is a social group in ground truth data, reporting {p, q, g} or reporting {p, q, r, s} will not be counted as true positive. Instead each of these instances will be counted as a false positive.

First, we measured the run-time performance of each method on two test datasets, 0424 and 0508. Fig. 4 shows the run-time of each method and dataset pair. Approximating with Convoys take the least amount of time while approximating with Flocks took the longest to report the results because flocks of single group can be detected multiple times in a time-window, but a group can form a single convoy for each time-window (by its definition). Our framework takes longer than approximating with convoys since, for each potential social group outputted from MPSG step, ISG step needs to compute its features and identify if it is a social group. Feature computation time dominates the ISG step.

Next, we measured the accuracy performance of the three methods. Analysis of the results for the datasets, 0424 and 0508, we learned that that the proposed framework gives higher accuracy (found 230/326 and 282/384 social groups for 0424 and 0508 datasets) than Convoys (found 221 and 267) and Flocks (found 67 and 78) as the approximating methods report many false positives (126 and 161 false positives were reported by Convoy while our framework reports only 71 and 91). Flock method performs the worst as also misses many of the social groups (259 and 306 false negatives from Flock method as compared to our method's 96 and 179 false negatives).

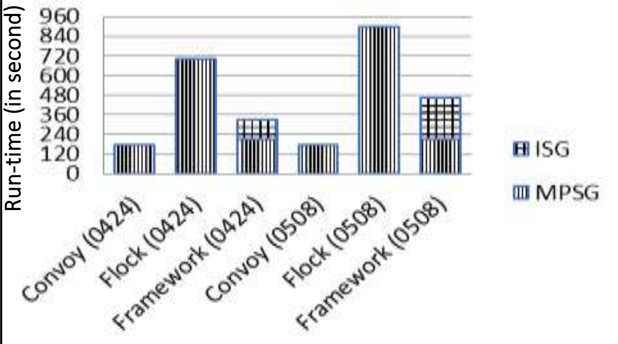


Figure 4. Run-time of the methods on each data

Table 4 compares the accuracy of the three methods in comparison in term of commonly used metrics, Precision, Recall, AUC (Area Under the Curve of ROC) and F1. Our proposed Framework yields far better AUC than Convoy and Flock, which performs worse than random chance due to the false positives. F1 values reflect a similar trend.

Table 4. Summary of the Datasets

Dataset	Method	Precision	Recall	AUC	F1
0424	Convoy	0.64	0.67	0.33	0.65
	Flock	0.32	0.21	0.10	0.25
	Framework	0.76	0.71	0.73	0.74
0508	Convoy	0.62	0.70	0.35	0.66
	Flock	0.29	0.20	0.10	0.24
	Framework	0.76	0.61	0.68	0.76

From the experiments, we concluded that approximating social groups using traditional convoy methods (Flock) cannot produce accurate results. Both Flock and Convoy methods cannot produce results better than random chances (ROC AUC of 0.1x and 0.3x respectively). Our proposed framework can produce more accurate results (ROC AUC of 0.7x) than simple convoy mining methods at a reasonable run-time, which even is faster than Flock by several minutes.

7. Conclusion

In this paper, we studied how to extract social groups from human mobility datasets. Social groups do not necessarily translate to a convoy. We reported that approximating social groups using traditional convoys does not yield a satisfactory result in real-life datasets. We proposed a two-step framework to discover genuine social groups from streaming human mobility data in real-time. The first step of our framework narrows down the search space while the second step employs machine learning techniques to model and correctly output social groups accurately. In experiments using real-life datasets, our proposed framework outperforms other approaches.

8. References

[1] Dražen Brščić , T. Kanda, T. Ikeda and T. Miyashita, "Person Tracking in Large Public Spaces Using 3-D Range Sensors," in IEEE Transactions on Human-Machine Systems, vol. 43, no. 6, pp. 522-534, Nov. 2013.

[2] H.H. Aung, L. Guo, K-L. Tan. "Mining Sub-trajectory Cliques to Find Frequent Routes." In Proc. Advances in Spatial and Temporal Databases. SSTD 2013. Lecture Notes in Computer Science, vol 8098. Pp. 92-109.

[3] F. Zanlungo, T. Ikeda and T. Kanda. "Potential for the dynamics of pedestrians in a socially interacting group." Physical review. E, Statistical, nonlinear, and soft matter physics vol 89 1 (2014). pp 012811.

[4] J. Gudmundsson and M. Kreveld. "Computing longest duration flocks in trajectory data." In Proc. the 14th annual ACM international symposium on Advances in geographic information systems (GIS '06). pp 35-42.

[5] M. Vieira, P. Bakalov, and V. Tsotras. "On-line discovery of flock patterns in spatio-temporal data." In Proc. the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09). pp 286-295.

[6] H. Jeung, H. T. Shen and X. Zhou, "Convoy Queries in Spatio-Temporal Databases," In Proc. IEEE 24th International Conference on Data Engineering, 2008, pp. 1457-1459.

[7] H.H. Aung and KL. Tan. "Discovery of Evolving Convoys." In: Proc. Scientific and Statistical Database Management. SSDBM 2010. pp. 196–213.

[8] M. Ester, H-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise." In Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD'96). pp 226-231.

[9] D. Mccaffrey, G. Ridgeway and A. Morral. "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies". Psychological methods. 9. pp 403-25.

[10] Jerome H. Friedman, "Stochastic gradient boosting", Computational Statistics and Data Analysis, Vol 38. Issue 4, 2002, pp 367-378.

[11] D. Brscic, T. Kanda, T. Ikeda, T. Miyashita, "Person position and body direction tracking in large public spaces using 3D range sensors", IEEE Trans. on Human-Machine Systems, Vol. 43, No. 6, pp. 522-534, 2013.

Performance Analysis of Parallel Clustering on Spark Computing Platform

Nway Yu Aung
University of Information Technology
nwayuaung@uit.edu.mm

Aye Chan Mon
University of Information Technology
ayechanmon@uit.edu.mm

Swe Zin Hlaing
University of Information Technology
swezin@uit.edu.mm

Abstract

In the area of information and technology, data is generated from a plethora of sources such as social media, internet of things, multimedia, sensor networks. Clustering is an essential data mining tool for analyzing this valuable information.Clustering algorithms are generally classified as a hierarchical and partitioning algorithm. This paper interested in partitioning algorithms. There are two kinds of partitioning algorithm, mean-based and medoids-based. The paper focuses on medoids-based because of medoids less influence by outliers or other extreme values than mean. But, one of the main issues of partitioning algorithm cannot handle large volume of data in case of the poor cluster quality and higher execution time.The objective of theresearchis to solve these two issues.To improve clustering quality, this paper appliesswarm intelligence optimization algorithm on the partition clustering algorithm. And then, this paper expects to reduce execution time for clustering large volume of data by using Spark framework.

Keywords- Clustering, Partitioning algorithm, Bat algorithm, Apache Spark

1. Introduction

Data mining is defined as a process used to extract usable data from a larger set of any raw data. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions [15]. Data mining involves effective data collection and warehousing as well as computer processing. Data mining is also known as Knowledge Discovery in Data (KDD). Clustering means grouping the objects based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar(or related) to one another and different from (or unrelated to) the objects in other groups [16].

It is the main task of data mining and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, Bioinformatics, data compression, and computer graphics. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

Clustering algorithms are generally classified as hierarchical and partitioning algorithms. This paper is interested in partitioning algorithms. K clusters found by a partitioning method are of higher quality than the K clusters produced by hierarchical method [13]. The main objective of the partition clustering algorithm is to divide the data point into K partitions. Each partition will reflect one cluster. The technique of partition depends upon certain objective function. There are mainly four types of partitioning algorithm includes as K-Mean Algorithm, K-Medoid Algorithm (i.e. Partition Around Medoid-PAM), CLARA and CLARANS.

In K-Mean clustering algorithm, a cluster is represented by its centroid, which is usually the mean of points within a cluster. Other algorithms are medoids based algorithm. Medoids algorithms select k-medoids initially and then swap the medoids object with non-medoids thereby improving the quality of cluster. Medoids based clustering is more robust than mean-based [14]. Because, medoids less influence by outliers or other extreme values than mean. So, these paper pursuit of medoids-based clustering algorithms. Among them, PAM algorithm is less sensitive to outliers than other partitioning algorithms and easy to implement. Also, traditional medoids-based clustering algorithms cannot handle large amounts of data. The weakness of these algorithms is the large volumes of data,the less effectiveness and efficiency [1]. So, this paper is proposed with the new PAM algorithm for solving these issues. PAM will combine swarm optimization algorithm for improving the cluster quality. For solving the execution time problem, this system will be worked on Spark parallel computing platform.

The paper is structured as follows: Section 2 presents related works. Section 3 describes the theory background. Section 4 discussesthe proposed system and Section 5 provides the results of the experiment.

2. Related Works

In the recent years, many clustering algorithms have been proposed. [1] proposed a new algorithm for k-medoids clustering and test several methods for selecting initial medoids. The proposed algorithm calculates the distance matrix once and uses it for finding new medoids at every iterative step. And then, compare with the results of another algorithm. That proposed method has better performance than k-means clustering and significantly reduced computation time. Partition Around Medoid has been used for clustering of face data [2]. The results are increased robustness to noise and outliers in comparison to other clustering methods.

[3] proposed optimized big data k-mean using MapReduce in which they claimed to counter the iteration dependence of MapReduce jobs. They used a sequence of three MapReduce jobs for the purpose. In their approach sampling technique is used in the first MR job and in the final MR job the data set is mapped to centroids using the Voronoi diagram. Other clustering algorithms like density based have also been implemented in distributed platforms. Work done in [4][5] implemented the DBSCAN algorithm in MapReduce. [6] modified k-means to deal with large scale heterogeneous data sets. To the best of our knowledge of all the versions of k-means presented so far in MR, number of clusters need to be specified before the start of the algorithm. Also, for getting the optimal number of clusters multiple runs might be required.

Yang [9] developed Bat algorithm in 2010. It provided a timely review of the bat algorithm and its new variants. It reviewed and summarized wide range of diverse applications and case studies. And then, further researchtopics are discussed. Xin-She Yang [10] proposes a new Meta heuristic method. To combine the advantages of an existing algorithm into the new bat algorithm. After a detailed formulation and explanation of its implementation, they compare the proposed algorithm with other existing algorithms, including genetic algorithms and particle swarm optimization (PSO).

3. Theory Background

3.1. Partition Around Medoid (PAM)

Partition Around Mediods (PAM) is developed by Kaufman and Rousseuw in 1987. It is based on the classical partitioning process of clustering. The algorithm selects k-medoids initially and then swaps the medoids object with non-medoids, thereby improving the quality of the cluster. This method is comparatively robust than K-Mean particularly in the context of noise or outlier. Medoids can be defined as that object of a cluster, instead of taking the mean value of the object in a cluster according to reference point. K-Medoids can find the most centrally located point in the given dataset. PAM algorithm is described in the following:

```
PAM Algorithm:
Input:
    • K:The number of clusters
    • D: A data set containing n objects
Output:
    • A set of K clusters
Method:
1. The algorithm begins with arbitrary selection of the K objects as medoids points out of n data points(n>K).
2. After selection of the K-medoids points, associate each data object in the given data set to most similar medoids.
3. Randomly select non-medoids object O.
4. Compute total cost S of swapping initial medoids object O.
5. If S>0, swap initial medoids with the new one.
6. Repeat steps until there is no change in the medoids.
```

In the PAM algorithm, the initial medoids is chosen by random. The result of such algorithm is highly depending on the initial choice of medoids and in the process of optimizing the objective function it may get stuck in local optima. To overcome these limitations, the nature inspired techniques have been proposed. These nature inspired techniques are decentralized and self-organized in behavior.The combination of PAM clustering and the nature inspired techniques makes as a hybrid approach which is very useful in the data mining.

3.2. Swarm intelligence algorithms

Swarm intelligence (SI) is one of the categories of nature-inspired problem-solving techniques. Swarm uses their environment and resources more efficiently by collective intelligence. There are many nature inspired techniques like artificial bee colony algorithm (ABC), particle swarm Optimization (PSO), ant colony optimization (ACO), bat algorithm (BA), firefly algorithm and glowworm swarm optimization (GSO). Among them, the proposed system chooses the bat algorithm because ofthis algorithm possesses the advantage of simplicity and additionally flexibility.

3.2.1. Bat algorithm. Bat algorithm is one of swarm intelligence-based algorithm which is worked on the echolocation of bats. This algorithm was developed by Xin-She Yang in 2010. Bats algorithm is based on the echolocation behavior of bats. They can find their prey or food and also, they can know the different type of insects even in a complete darkness. These bats use a type of sonar namely as echolocation. They emit a loud sound pulse and detect an echo that is coming back from their surrounding objects.

Their pulse variation in properties and will be depend on the species. Their loudness is also varied. When they are searching for their prey, their loudness is loudest if they are far away from the prey and they will become slow when they are nearer to the prey. For the emission and detection of echo which are generated by them, they use time delay. And this time delay is between their two ears and the loudness variation of echoes. The propose system based on the echolocation behavior of bats to know the initial value to overcome the partition around medoids issue.

Bat Algorithm:

Step 1 : Set defined appropriate constant parameters

Step 2 :Bat movement is calculated by the three equations

(i) Update the frequency

$$f_i = f_{min} + (f_{max} - f_{min})\beta$$

(ii) Update the velocity

$$v_i^{t+1} = v_i^t + (v_i^{t-1} - x_*)f_i$$

(iii) Calculate the next position

$$x_i^{t+1} = x_i^t + v_i^t$$

Step 3 : The fitness function defined to be the total dissimilarity between every object and the medoids of its cluster

The traditional bat algorithm is modified to the clustering process by randomly assigning k-clusters to each of the N bats. In the next stage, fitness of medoids in each bat is computed. The data items or objects are placed in proper cluster that is based on the fitness value of medoids in a bat. In successive generations, the new solution is generated by adjusting the frequency, updating the velocity and creating new medoids values. For each bat, the best solution is selected among a set of the best solutions from the other bats. To accept new solution, the frequencyis increased and reduced loudness is considered. Based on the newly selected solution, clusters are reassigned for medoids update assignment.

In the proposed system, we combine the two algorithms for the purpose of achieving better performance in data clustering. Also, this system intends to cluster large volume of data.Large-scale data has turned to parallel and distributed processing by providing advanced mechanisms. How to wisely use existing parallel frameworks with large-scale data becomes the biggest challenge.

Hadoop MapReduce and Spark are most popular frameworks in many open-source parallel frameworks. Spark can do it in-memory, while Hadoop MapReduce has to read from and write to a disk. As a result, the speed of processing differs significantly. Spark may be up to 100 times faster. If the tasks process data again and again – Spark defeats Hadoop MapReduce. Spark's Resilient Distributed Datasets (RDDs) enable multiple map operations in memory, while Hadoop MapReduce has to write interim results to a disk. The main factor of the proposed partitioning clustering algorithm is iterative nature. Apache Spark is more suitable for iterative algorithms. So, the proposed system chooses this framework.

3.3. Apache Spark

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. Spark has several advantages compared to other big data and MapReduce technologies like Hadoop and Storm. Spark supports machine learning, SQL queries, graph data processing and streaming data for analysis. Spark supports languages such as Java and Python, and it is implemented in Scala, and it runs on the Java Virtual Machine (JVM).

Spark consists of cluster manager, driver program (spark context), executor or worker and HDFS. In spark, a Driver program is considered as the main program. Spark Context is for the coordination of the applications which run on clusters as a set of processes. The processes used for applications are assigned uniquely, i.e. they all have their processes and due to this task run in multiple threads, and they must connect to worker nodes. These worker nodes run computations and store the data. Programming is written in Java or Python language which is sent to the executor and it runs the tasks. The two main key concepts used in Apache Spark are Resilient Distributed Datasets (RDD) and Directed Acyclic Graph (DAG).

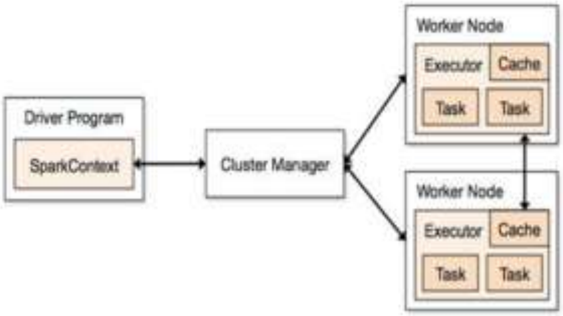


Figure 1: Apache Spark

3.3.1. Resilient Distributed Dataset (RDD). Resilient Distributed Datasets work as a collection of elements which are operated in parallel. In the distributed file system Spark runs on Hadoop cluster, and RDD is created from files in the format of text or sequence files. RDD is used for reading the objects in the collection and when some partition is lost, it can be rebuilt because RDDs are distributed across a set of machines.

4. Proposed System

The research aim is to improve the performance efficiency and effectiveness of the traditionalpartitioning clustering algorithm for handling the large amount of data. This proposed system has two portions. First, the system mentions the execution time of clustering. And then,the system discusses the cluster quality.This paper only focused on execution time.

4.1. Pre-processing

The first portiondiscussed on execution time of two algorithms. This step tested on parallel PAM and Bat algorithms by using Apache Spark.

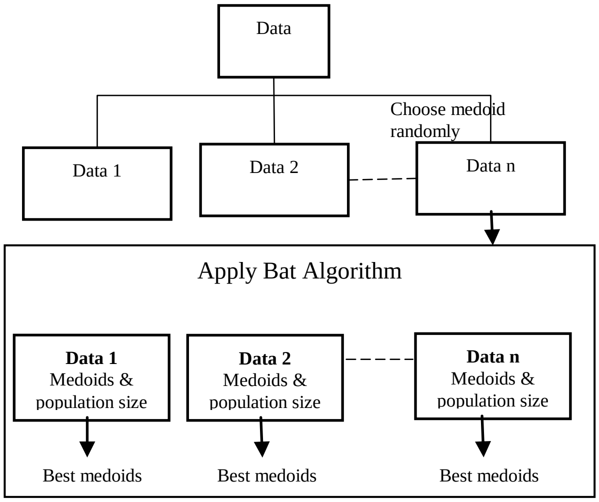


Figure 2: Parallel Bat algorithm

Figure 2 show the parallel bat algorithm. The data is distributed on the number of workers of apache spark. And then, we apply the bat algorithm for choosing the best medoids. PAM algorithm also work parallel on applying the apache spark as shown in the figure 3.

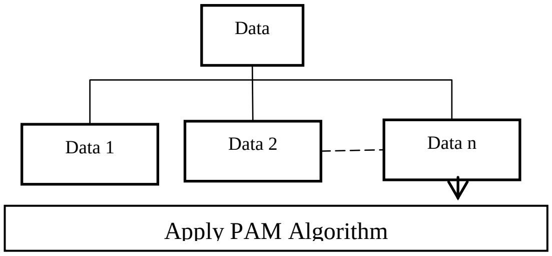


Figure 3: Parallel PAM algorithm

4.2. Bat-PAM Hybrid Algorithm

The secondpart of research work concerns about the quality of cluster results**.** Traditional PAM chooses medoids randomly. This is mainly affected of cluster quality. Bat algorithm is choosing the best of medoids by assigning the population of bats around the medoids. Best medoids that have chosen by Bat is parameters in PAM algorithm. The proposed system architecture shows in figure 4. But, this paper doesn't mention this portion.

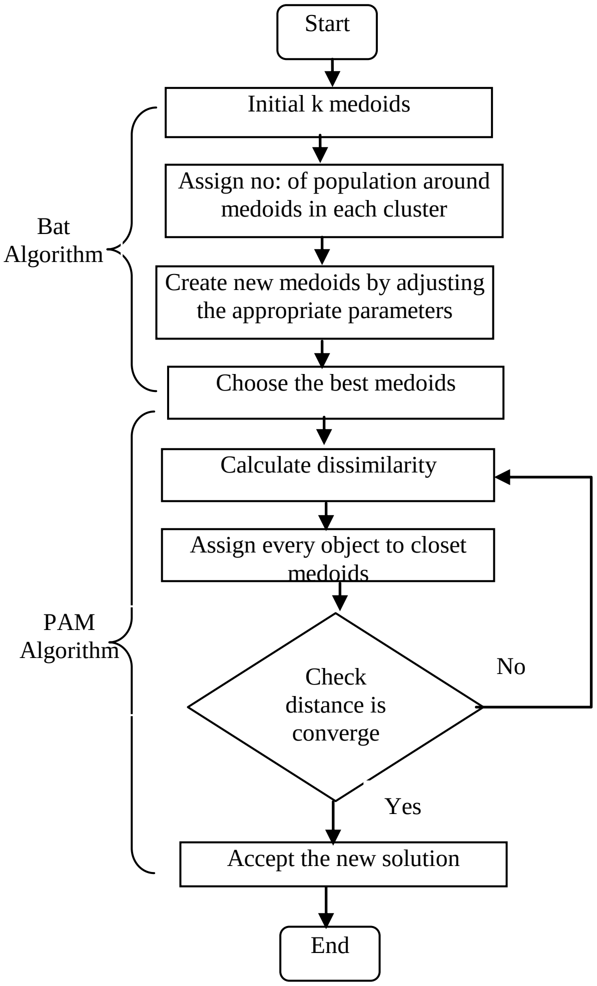

Start
Initial k medoids
Assign no: of population around medoids in each cluster
Create new medoids by adjusting the appropriate parameters
Choose the best medoids
Bat Algorithm
Calculate dissimilarity
Assign every object to closet medoids
Check distance is converge
No
Yes
PAM Algorithm
Accept the new solution
End

Figure 4: The Proposed System architecture

5. Experimental Result

The system tests the performance efficiency of parallelization technique of two algorithms. The system implemented on a personal computer with an Intel (R) Core (TM) i7-4770 CPU (3.40GHz) with8GB RAM.The system uses the sample Census dataset about 500MB. This dataset contains 1,000,000 records and 68 attributes.

First, we test the execution time of the dataset in sequentially. Next, we implemented the standalone

Apache Spark. Spark Standalone deployment means Spark occupies the place on top of HDFS (Hadoop DistributedFile System) and space is allocated for HDFS, explicitly.

Figure 5 shows the execution time of traditional PAM and PAM on Apache Spark. The input data is stored in Hadoop DistributedFile System (HDFS). So, we first need to load these inputdata into RDDs, where the data is split anddistributed across all nodes.The results show that parallel technique with spark framework significantly outperforms.

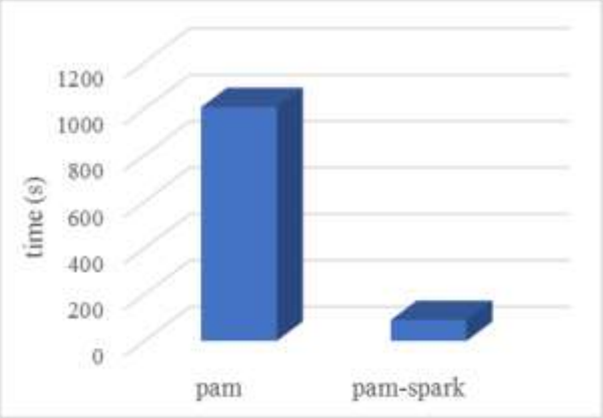
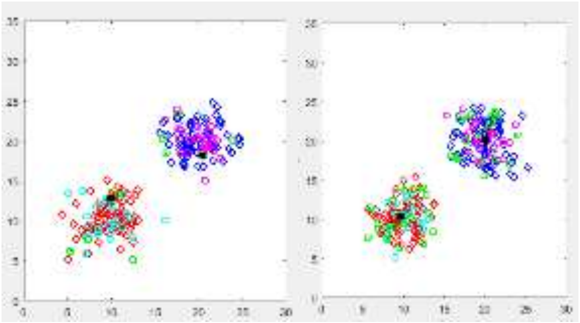


Figure 5: Comparison of execution time

The performance of apache spark is depend on the executor memory and use of cores, and the number of worker nodes. So, we consideredthese facts. At that point, the size of dataset is a factor that we need to consider. In the system, we configure one master and two worker nodes. Thiscan handle the current dataset size. In tend to use larger size of the data we will adjust this.

And then, we discussed the bat algorithm for clustering sequentially. This system simulated the number of bat population from n=5 to n=20 as shown in figure 6.



(a) n=5 (b) n=20

Figure 6: Different numbers of populations (n) on dataset

We observe the execution time of different number of populations on the dataset.

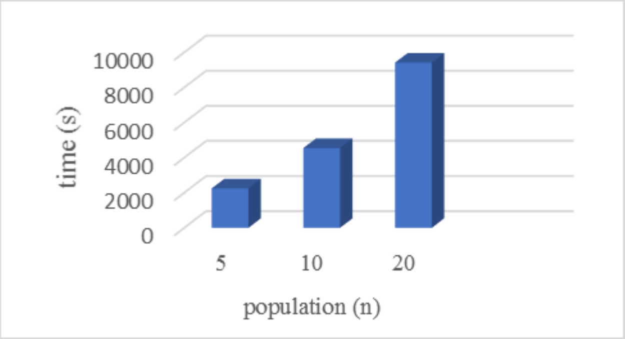


Figure 7: Comparison the runtime on the different number of bat populations

The experimental results show that the lower population size of bats causes quickly converge. Higher population size may cause slow convergence.

6. Conclusion and Future Works

Clustering is always confronted with questions such as an unstable clustering result, low executive efficiency. For solving the higher execution time problem, the system used the Apache Spark Parallel Computing Platform. Further development of this system,the Partition around medoids algorithm will combineBat algorithm to get the better cluster quality. Also, we will test the real large volume of the Myanmar census dataset. And then, we will compare the cluster quality and execution time with other well-known clustering algorithms.

7. References

[1] Hae-Sang Park, Chi-Hyuck Jun, “A simple and fast algorithm for K-medoids clustering”, Expert System with applications, ELSEVIER, 2009.

[2] Aruna Bhat, “K-medoids clustering using Partitioning arund medoids for performing face recognition”, International Journal of Soft Computing, Mathematics and Control (IJSCMC), Vol.3, No.3, August 2014.

[3] Cui, Xiaoli and Zhu, Pingfei and Yang, Xin and Li, Keqiu and Ji, Changqing, “Optimized big data K-means clustering using MapReduce”, The Journal of Supercomputing, 2014.

[4] Kim, Younghoon, Kyuseok Shim, Min-Soeng Kim, and June Sup Lee, “DBCURE-MR: an efficient density-

based clustering algorithm for large data using MapReduce”, Information Systems 42, 2014.

[5] He, Yaobin, te al, “Mr-dbscan: an efficient parallel density-based clustering algorithm using mapreduce”, Parallel and Distributed Systems (ICPADS), 2011 IEEE 17th International Conference on IEEE (2011).

[6] Cai, X., Nie, F., and Huang, H.., “Multi-view k-means clustering on big data”, In proceeding of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press (2013).

[7] Ankita Sinha, Prasanta K.Jana, “A Novel K-Means based Clustering Algorithm for Big Data”, International Conference on Advances in Computing, Communications and Informatics, 2016.

[8] Yasmine Aboubi, Habiba Drias, Nadjet Kamel, “BAT-inspired algorithm for Clustering LARge Applications”, ELSEVIER (2016).

[9] Xin-She Yang, “Bat Algorithm: Literature Review and Applications”, International Bio-Inspired Computation, 2013.

[10] Xin-She Yang, “A New Metaheuristic Bat-Inspired Algorithm”, Computational Intelligence, Springer (2010).

[11] Tsai, C.W., Huang, W.C., and Chiang, M.C. “Recent development of metaheuristics for clustering”, In Mobile, Ubiquitous, and Intelligent Computing (2014).

[12] Tsutomu, S., Fumihiko, Y., and Yoshiaki, T., “A new algorithm based on metaheuristics for data clustering”, SCIENCE AISSN (2010).

[13] Yasmine Aboubi, Habiba Drias, Nadjet Kamel, “BAT-inspired algorithm for Clustering Large Applications”, 8th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2016: Troyes, France, 28-30 June 2016.

[14] A.Dharmarajan, T.Velmurugan, “Efficiency of k-Means and k-Medoids Clustering Algorithms using Lung Cancer Dataset”, International journal of Data Mining Techniques and Applications, 2016.

[15]https://economictimes.indiatimes.com/definition/data-mining.

[16]https://www.slideshare.net/archnaswaminathan/cdm.

Community Detection in Social Network Using Artificial Bee Colony with Genetic Operator

Thet Thet Aung
University of Computer Studies, Yangon
thetthetaung@ucsy.edu.mm

Abstract

Community detection (CD) plays an important role in analyzing social network features and helping to find out valuable hidden information. Many research algorithms have been proposed to find the best community in the network. But it has many challenges such as scalability and time complexity. This paper proposes a new algorithm, Artificial Bee Colony Algorithm with Genetic Operator (ABCGO) that combines crossover and mutation operators with Artificial Bee Colony algorithm. This paper takes modularity Q as objective function. Compared with five state-of-art algorithms, results on real world networks reflect the effectiveness of ABCGO.

Keywords- Social Network, Community Detection, Artificial Bee Colony, Modularity

1. Introduction

Social network is a theoretical abstraction, useful in social sciences to study the relationships between individuals, groups, or organizations. The connected items are persons or organizations and the ties are interaction or communication between pairs of actor. Real world social networks are usually found to divide naturally into small communities. Generally, a community in a social network is a set of nodes that are densely connected internally, but loosely connected to the rest of this network. In recent years, community detection in a network has become one of the main topics of fields, such as biology, computer science, physics, and applied mathematics. Great applications of CD are to detect suspicious event in telecommunication networks, link prediction, refactoring the software package, recommendation and containment of virus and warm online, recently for criminal detection in large network.

Networks could be modeled as graphs, where nodes represent the objects and edges represent the interactions among these objects. Communities play special roles in the structure-function relationship, and detecting communities can be a way to identify substructures that may correspond to important functions in the network.

With the development of technology, more and more data can be gotten from all kind of social networks, such as facebook, twitter, others. The big data processing is one challenge in many research fields because the efficiency of previous algorithms cannot be guaranteed with the increasing of the data and the network. So, the effective algorithm for community detection is also needed in Social Network.

Network community detection can be viewed as an optimization problem. Due to their inherent complexity, these problems often cannot be well solved by traditional optimization methods. For this reason, nature inspire algorithms have been adopted as a major tool for dealing with community detection problems. In this work, Artificial Bee Colony with Genetic operator (ABCGO) has been used as an effective optimization technique to solve the community detection problems. Modularity is used to measure the community result because it is one of the popular fitness measures for community structures.

The paper is organized as follow. Related works is presented in section 2. Theory background is in section3 and proposed approach is detailed in section 4. Experimental results are shown in section 5. Finally it is concluded in session 6.

2. Related Works

Community detection is similar to a graph partitioning problem. Most of the graph partition methods are based on optimizing a quality function. Hierarchical clustering techniques depend on similar measure between vertices to form clusters. It is classified under two categories, Agglomerative algorithm and Divisive algorithm. In agglomerative algorithm, starts from vertices as separate communities and ends up with a graph as unique community. Divisive algorithm takes the opposite direction of agglomerative. It starts from a graph as one cluster and ends up with clusters containing similar vertices by removing edges in where the authors use betweenness measure to remove iteratively edges from the network to split it into communities [1]. One community detection algorithms proposed is the Girvan-Newman (GN) algorithm, which introduces a divisive method that iteratively removes the edge with the greatest betweenness value [1].

Maps of random walks (Infomap) proposed by Rosvall and Bergstrom [2]. This algorithm is a flow-based and information theoretic clustering approach. It uses a random walk as a proxy for information flow on a network and minimizes a map equation, which measures the description length of a random walker, over all the network clusters to reveal its community structure. Infomap aims to finding a clustering which generates the most compressed description length of the random walks on the network. Optimization-based methods have been considered as the main category. Optimization method can be divided into two categories; single-objective and multi-objective optimization. Both are proved to be efficient and effective for optimization problem.

Hafez, et al. proposed in artificial bee colony algorithm this employs three types of bees to solve the community detection problem and show how the algorithm performance is directly influenced by the use of different community measures [3].

Gema et.al proposed evolutionary clustering algorithm for community detection using graph based information. In their approach, genetic algorithm with fitness that combines different measure of network topology is used for clustering. Binary encoding was used where binary 1 used to denote the node belong to the community [4]

Discrete Particle Swarm Optimization for community detection problem is proposed by Zhou et.al. Modularity density function have been used for objective function in their approach and particle status updating for discrete PSO have been proposed for the community detection problem [5].

Youcef Belkhiri et.al proposed bee swarm optimization for community detection in complex network. This proposed algorithm takes modularity as objective function and k number of bee to create a search area. The algorithm starts with initial solution called reference solution and the taboo list to avoid cycles during the research process [6].

In this paper, genetic operator based ABC with label based encoding will be used. Modularity will also be used for the objective function of the proposed system.

3. Background

In this section includes the definition of social network, community detection, artificial bee colony algorithm and other definition.

3.1. Social Network

In a social network, G(V,E), where V is a set of nodes and E is the edges between the nodes, a community is a group of nodes with tightly connected edges with each other. The nodes in a community show similar characteristics. For example, in social network, people in a community show similar interest to a trend in a community, for example, buying the same products in online marketing.

3.2. Community Detection

A community (also called cluster or module) in a network is a group of vertices having a high density of edges within them and a lower density of edges between groups [7]. Community Detection is the process through which nodes in networks are clustered based on the connection between them. Nodes in community are densely connected and are sparsely connected to other communities.

The purposes of community detection are to understand the interaction between actors, visualize and navigate large network and forming the basis of other tasks such as data mining. By identifying community structure in network can provide knowledge about how network function and topology affect each other.

Community structure is also named as cover of community. It is a set of communities present in network. It is represented as C={c_1,c_2,c_3,c_4,…,c_j}. C is the communities structure and c_1, c_2, c_3, c_j are communities. Figure 1, There are two communities c_1={1,2,3,4} and c_2={5,6,7,8,9}.

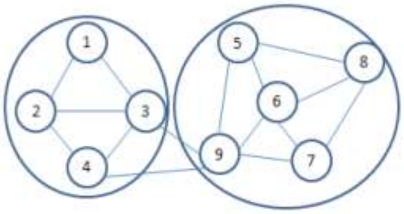


Figure1. Example of communities present in a network

3.3. Artificial Bee Colony Algorithm

Artificial bee colony is one of the most recently defined algorithms by Dervis Karaboga in 2005[8], motivated by the intelligent behavior of honey bees. ABC as an optimization tool provides a population based search procedure in which individuals called foods positions are modified by the artificial bees with time and the bee’s aim is to discover the places of food sources with high nectar amount and finally the one with the highest nectar. It contains three groups: scouts, onlookers, and employed bees. Onlookers and employed bees carry out the exploitation process in the search space. Scouts control the exploration process.

The general algorithmic structure of the ABC optimization approach is given as follows:

```
Initialization Phase
REPEAT
        Employed Bees Phase
        Onlooker Bees Phase
        Scout Bees Phase
        Memorize the best solution achieved so far
    UNTIL (Cycle = Maximum Cycle Number or a Maximum CPU time)
```

In the initialization phase, the population of food sources (solutions) is initialized by artificial scout bees and control parameters are set.

In the employed bees phase, artificial employed bees search for new food sources having more nectar within the neighborhood of the food source in their memory. They find a neighbor food source and then evaluate its fitness. After producing the new food source, its fitness is calculated and a greedy selection is applied between it and its parent. After that, employed bees share their food source information with onlooker bees waiting in the hive by dancing on the dancing area.

In the onlooker bees' phase, artificial onlooker bees probabilistically choose their food sources depending on the information provided by the employed bees. For this purpose, a fitness based selection technique can be used, such as the roulette wheel selection method. After a food source for an onlooker bee is probabilistically chosen, a neighborhood source is determined, and its fitness value is computed. As in the employed bees phase, a greedy selection is applied between two sources.

In the scout bees' phase, employed bees whose solutions cannot be improved through a predetermined number of trials, called "limit", become scouts and their solutions are abandoned. Then, the scouts start to search for new solutions, randomly. Hence, those sources which are initially poor or have been made poor by exploitation are abandoned and negative feedback behavior arises to balance the positive feedback.

These three steps are repeated until a termination criteria is satisfied, for example a maximum cycle number or a maximum CPU time [9]. This paper used the backbone of Artificial Bee Colony and combined with genetic operators.

3.4. Objective Function

Many objective functions for community detection that can capture the intuition of communities have been introduced from different research fields. Quality functions can be used when there is no ground truth for the communities to assess the quality of detected communities. Some of the objective functions such as conductance, expansion, cut ratio, community score and modularity that are already widely used in community detection literatures or can be used for community detection [10]. In this paper, modularity is used to measure the quality of result community.

Newman and Girvan first defined a measure known as 'modularity' to judge the quality of partitions or communities formed. The modularity measure proposed by them has been widely accepted and used by researchers to gauge the goodness of the modules obtained from the community detection algorithms with high modularity corresponding to a better community structure. Network modularity function, also called Q-function, is widely used to quantitatively evaluate the community partition of complex networks.

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad 1$$

In equation 1, where m is the number of edges in network, k_i and k_j are the degree of nodes i and j respectably. A_{ij} is Adjacency Matrix. c_i and c_j denotes the communities of nodes i and j respectably. Function δ (c_i $_,c_j$) is 1 if i and j are in the same community, otherwise it is 0. The value of Q is between -1 and 1 [11].Objective function plays an important role in the optimization process that leads to good solution. There are many objective function have been proposed to capture the intuition of communities.

4. Proposed Algorithm

ABC algorithm is primarily widely used for real optimization problem. For the CD problem, discrete combinational operators are needed. So ABC with genetic operators is inspired from the genetic algorithm for CD that enables us to use the ABC algorithm for combinational optimization problem.

4.1 Artificial Bee Colony with Genetic Operator

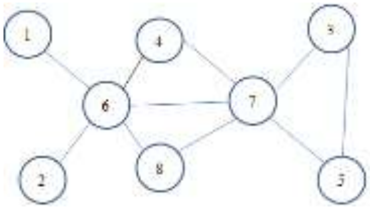
ABC is well known for proven result in numerical optimization problems. Theoretically community detecting problem is NP-hard problem and thus people inclined to choose heuristic algorithms based on objective optimization. Traditional ABC algorithm is suitable for solving real point optimization problem and not suitable for discrete problem like community detection. So, ABC algorithm is modified with genetic operators (crossover and mutation) for community detection. String-based representation is used in the ABC algorithm in which each locus value represents the community index in which it belongs to. Modularity is used as the fitness function of the ABC. In ABC algorithm, exploitation is performed by employed bees and onlookers, while scout bees do the exploration. Employee bee and onlooker bee exploit the food source to create new better food source or solution. In the proposed algorithm, crossover and mutation is used for creating new food source from neighbor.

For the crossover operation, two food source is selected based on their fitness probabilities, one point crossover is performed, from the resulting two food source, greedy selection is applied. Mutation performs the exploration function of the algorithm, in which locus value of food source is randomly mutate to the community index of one of their neighbors allowing the algorithm to explore the search space that have been unexplored.

4.2. Encoded form

According to the nature of community detection problem; a solution is partitioning of nodes V of network G. Each partition contains similar nodes and represents a community. Algorithm starts with initial population creation. An integer array arr is used for data representation of community detection problem. Array store community identifier (CID) of nodes, that arr_i is the community identifier of the node i. The array has n elements and is called as *individual* food source in ABC or *chromosome* in genetic algorithm [12]. There are number of individuals holding different community configuration information in the population. Each individual produces a possible solution.

A possible solution is defined by the number of communities and by the distribution of the nodes in these communities. To represent the assignment of each node in the network for each community, string representation is used. In representation, each locus value indicates the community number the gene belongs to. Figure2 (a) shows a network of 8 nodes to be partitioned: (b) is an initial bee source of this network and the food source of this network represents the reference solution, where node 1, 2,4,6,8 are in community number 1 and nodes 7,3,5 form community number 2.The network is divided into 2 communities and (c) is the community structure of the proposed bees source.



(a)

Food source

Node	1	2	3	4	5	6	7	8
CID	1	1	2	1	2	1	2	1

(b)

(c)

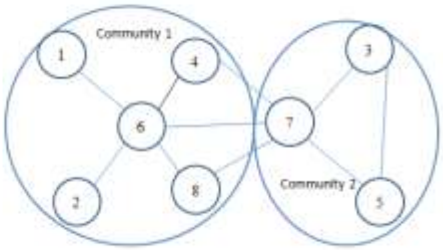


Figure 2: Encoding initial bee strategy for a network

4.3. Genetic Operator

Traditional crossover and mutation of the ABC algorithm need to modify to work with the proposed solution. Crossover uses one point crossover. Pick up a crossover point randomly. Following example shows using graph in figure 2 (a). Choose two individual for crossover.

Food source 1

1	2	3	4	5	6	7	8
1	1	2	1	2	1	1	1

Food Source 2

1	2	3	4	5	6	7	8
1	1	2	2	2	1	2	1

Eg. Crossover point=4

Select gene index with same community as crossover point depend on food source 1. In example, node 4's community id is 1. Then, choose other nodes that community id is the same with node 4.

Food source 1

1	2	3	4	5	6	7	8
1	1	2	1	2	1	1	1

Food Source 2

1	2	3	4	5	6	7	8
1	1	2	2	2	1	2	1

Crossover points [1,2,4,6,7, 8]

Exchange gene values in food source 1 with same gene indexes values in food source 2.

Food Source 1

1	2	3	4	5	6	7	8
1	1	2	2	2	1	2	1

Food Source2

1	2	3	4	5	6	7	8
1	1	2	1	2	1	1	1

Pick up solution with best modularity fitness value. Equation 1 is used in food source 1 and 2 to find better community quality. Food source1's fitness value is 0.195 and food source2's fitness value is 0.12.

For mutation

For each gene, pick up random probabilities, if the rand< mutation probabilities then go on mutation. Example showed, food source 2 gets better fitness value. For genes to be mutated, replace the community index with neighbor node community index. Choose food source 2 to mutate. Node 7 is neighbor with node 4, 6, 8 and 3, the community of node 4,6,8 is 1 and 3 is 2 respectively. It is replaced node 7's community number with 1. This process continues until Maximum Cycle Number is reached.

4.4. Objective function

Objective function plays an important role in the optimization process that leads to good solution. The problem of community detection in Social Network is NP-hard and requires a quality function in order to evaluate and discover good communities. In this work use the modularity Q defined in equation 1. When community group of network is gotten, this community quality is tested by using quality function. This quality function Q assigns a number of each community in a network and communities are ranked based on the score of Q. Good community has higher modularity value. The real network modularity value is generally between 0.3 and 0.7.

5. Experimental Result

In order to show the efficiency of proposed algorithm, choose four real networks used to run. They are Zachary Karate Club [13], Dolphin Network[14] Football Network[15] and Facebook Network [16].

Zachary Karate Club dataset contains friendship network between 34 members of Karate club at US University in 1970. It is one of the most widely used networks in CD. The relationships between members constitute the 78 edges of the network.

Dolphin network is based on the observations of the behavior of 62 dolphins over a period of seven years living in Doubtful Sound, New Zealand. It contains 62 dolphins as nodes and 159 connections as the edges in the network.

American College football game between American colleges during regular season fall 2000. It contains 115 teams as nodes and 613 edges.

Facebook Network dataset in SNAP consists of friends lists from facebook. Facebook data was collected from survey participants using facebok app. It contains 4039 nodes and 88234 edges.

Traditional community detection algorithm is tested on the four real dataset. Some algorithms have good modularity for small network but they can always not get good modularity for large scale graph. Researchers also use population based algorithm to find community detection. This paper uses other five traditional community algorithms [17-21] which are state-of-art method in community detection research.

ABCGO algorithm for community detection is also tested on the four real networks. The results of the algorithm and the other methods are showed in figure 3. Figure 3 illustrates a numeric quality comparison of proposed algorithm with other algorithms, x-axis represents different CD algorithms and y-axis is the modularity value for these datasets. The proposed algorithm gets suitable modularity values in Zachary Karate club, football and facebook datasets. In dolphin, their modularity results are small difference. ABCGO gets excellent result in football dataset because its nodes contain many link connections to the other nodes (densely connected network). The algorithm gets efficient result in large dataset such as football and facebook. The result community quality depends on modularity value. Some algorithms get suitable modularity result but the number of community is different with ground truth. ABCGO uses prior information (the number of community structure) which make the algorithm more targeted and improves accuracy of community detection

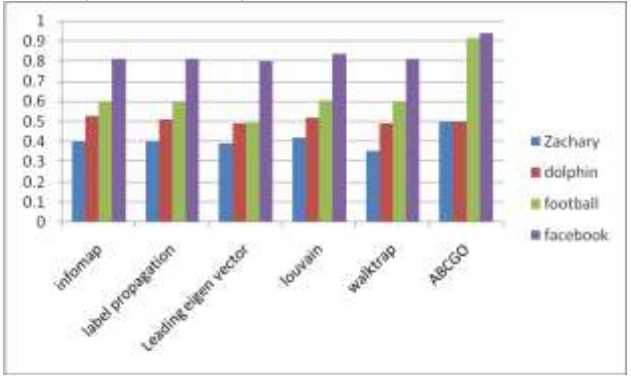


Figure 3. ABCGO on real networks and other algorithms' clustering Modularity quality comparison

.

For proposed algorithm, initial numbers of population and maximum cycle number are important. Population size was 50 and maximum cycle number is 100, crossover rate is 0.6 and mutation rate is 0.1 are used to test the algorithm.

6. Conclusion

Nature inspire algorithms can improved by adjusting balance between exploitation and exploration. This paper proposed a new approach for community detection in social networks called ABCGO. The approach is based on population algorithm to find the community structure and the merging of communities. The approach is constructed for undirected and un-weighted networks. Results obtained on real social network argue for the capacity of the approach to detect real communities. For large scale network dataset, the efficiency and scalability of a community detection algorithm is crucial for its popularity. Future development focuses on the scalability of community detection problem with parallel ABCGO and set some criteria for increasing the accuracy.

7. References

[1]. M. Girvan and M. E. J. Newman.”Community structure in social and biological networks.” Proceedings of the National Academy of Sciences, 99:7821–7826, 2002.

[2]. M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences of the United States of America, 105(4):1118–23, Jan. 2008.

[3]. Ahmed Ibrahem Hafez, HossamM.Zawbaa, Aboul Ella Hassanien, Aly A. Fahmy, “ Network Community detection using artificial bee colony swarm optimization”. http://scholar.cu.edu.eg/sites/default/files/abo/files/ibica2014_p27.pdf

[4]. Gema Bello-Orgaz , David Camacho, “Evolutionary clustering algorithm for community detection using graph-based information” ,2014 IEEE Congress on Evolutionary Computation (CEC) July 6-11, 2014, Beijing, China

[5]. Zhou, D.-Q ,Wang, X ,Cheng, S.-Y , Chen, Y. (2016). “Community detection algorithm via discrete PSO”. 38. 428-433. 10.3969/j.issn.1001-506X.2016.02.28.

[6]. YoucefBelkhiri, NadjetKamel, HabibaDrias, and SofianeYahiaoui, “ Bee Swarm Optimization For Community Detection in Complex Network”, Spinger International Publishing AG 2017.

[7]. Mehjabin Khatoon, W. Aisha Banu, " A survey on Community detection Methods in Social Networks", I.J. Education and Management Engineering , 2015,1,8-18, May in MECS.

[8]. D. Karaboga, “An idea based on honey bee swarm for numerical optimization”, Technical Report, TR06, ErciyesUniversity,Engineering Faculty, Computer Engineering Department, 2005.

[9]. DervisKaraboga, BeyzaGorkemli, CelalOzturk, NurhanKaraboga, “A comprehensive survey: artificial bee colony (ABC)algorithm and applications”, 11 March 2012.

[10]. Chuan Shi and Yanan Cai, Philip S. Yu, Zhenyu Yan, Bin Wu , "A Comparison of Objective Functions in Network Community Detection", International Conference on Data Mining Workshops © 2010 IEEE

[11]. Mingming Chen, Konstantin Kuzmin, Student Member, IEEE, and Boleslaw K. Szymanski, “Community Detection via Maximization of Modularity and Its Variants”, IEEE trans. Computation Social System, vol. 1(1):46-65, March 2014

[12]. Mursel Tasgin and Haluk Bingol, “Community Detection in Complex Networks using Genetic Algorithm”, 89.75.Fb, 89.20.Ff, 02.60.Gf

[13]. Zachary, W.W, “ An information flow model for conflict and fission in small group.”, J. Anthropol.Res.33,452-473(1977)

[14]. Lusseau, D, “ The emergent properties of a dolphin social network.” Proc.R.Soc.Lond. B Biol. Sci. 270(Suppl 2), S186-S188(2003)

[15]. Girvan,M., Newman, M.E.J, “ Community structure in social and biological networks.”, Proc. Nat. Acad. Sci. 99(12), 7821-7826 (2002)

[16]. “Stanford Large Network Dataset Collection”, http://snap.stanford.edu/data/index.html,

[17]. M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences of the United States of America, 105(4):1118–23, Jan. 2008.

[18]. Raghavan, U. N., Albert, R. & Kumara, S. “Near linear time algorithm to detect community structures in large-scale networks”.Physical Review E 76, 036106 (2007).

[19]. Newman, M. E. “Finding community structure in networks using the eigenvectors of matrices”. Physical Review E 74, 036104 (2006)

[20]. Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: “Fast unfolding of communities in large networks”. J. Stat. Mech. (2008) P10008

[21]. Pons, P. & Latapy, M. “Computing communities in large networks using random walks”.In Computer and Information Sciences-ISCIS 2005, 284–293 (Springer, 2005).

Database and Big Data Analytics

Graph-based Household Matching for Linking Census Data

Khin Su Mon Myint, Win Win Naing
University of Information Technology
Yangon, Myanmar
ksmonmyint@uit.edu.mm, winwinnaing@uit.edu.mm

Abstract

Historical censuses consist of individual facts about a community. It provides knowledge concerned with the nation's population. These data apply the reconstruction features of a specific period to trace their ancestors and families changes over time. Linking census data is a difficult task as common names, data quality and household changes over time. During the decades, a household may split multiple households due to marriage or move to another household. This paper proposes a graph-based approach to link households, which takes the relationship between household members. Using individual record linking results, the proposed method builds household graphs, so that the matches are determined by attribute similarity and records relationship similarity. According to the experimental results, the proposed method reaches an F-score of 0.974on Ireland Census data, outperforming all alternative methods being compared.

Keywords-Historical Censuses; Data Matching; Record Linkage; Household Linkage; Group Matching

1. Introduction

The population census data provide useful information on a specific region. They play an important role in analyzing for the social, economic, education and demographic aspects of a population [7, 14, and 10] in that region. These data can also be used for planning or reconstruction purposes in the country. Censuses are taken regularly by governments every ten years. These data allow us to understand populations and their different characteristics such as population size, age structure, household compositions, occupations, and other socio-demographic aspects [13].

Historical censuses contain specific information also gives the state of the nation and facilitate the construction aspects such as birth, death, education, occupation, etc. Linking record refers to the same households from several censuses that give across the decades. It is the process of observing records that refer to the same entities from different databases. These records will greatly enhance in value. The linked results have been allowed to trace varies in the characteristics of individual households over time.

Linked information improves not only retrieving of information, but also supporting new opportunities for improving the quality of the data. It can also help social scientists with the dynamic character of social, economic and demographic changes [8], which helps the reconstruction of the region.

Difficulties of historical census data linkage include poor data quality due to census data collection process. Importantly, the situation of individuals in a household may vary significantly between two censuses. For example, people are born and die, get married, change occupation, or moved home. As a result, linking individuals is not reliable, and many false matches are often generated.

Due to the benefits of historical census data linkage, there are a large amount of data available, automatic or semi-automatic linking methods have been developed by data mining researchers [14, 10,7, 5]. These methods treat historical census data linkage as a special case of record linkage, and apply string comparison methods to match individuals. Some researchers use classification algorithms to classify matches or non-matches and use group linking approach to link households based on the matched records [4].

Most of the researchers aim to find households with the majority of their members matched. However, during the ten year interval between two censuses, a household may split into multiple households due to marriage or move out to another household, or servants may change jobs. Most previous works in the census household linking problem can only be matched each individual in one household to one individual in another household. Then, previous historical census matching method couldn't support the household structure changes between the decades. Then, they have not taken the relationship between the individuals in the household. If the relationship information between household members can be considered in the linking model, the linking accuracy can be improved.

This paper proposes a graph-based approach for linking of historical census data using the relationship between the individuals in the household. This work considers not only each individual in one household to one individual in another household but also takes multiple household linking.

The main idea of graph-based approach is to match multiple household records and all of them treat records that are linked to each other as vertices and links between them as edges. So, the edges show the similarity between individual members. The proposed approach builds household graphs and the vertices correspond to each household member, the edges show the relationship between members. Record linkage is performed on household graphs, and then the linking results are improved by considering the relationships between the records.

The rest of the paper is arranged as follows. Section 2 introduces related works in record linking. Section 3introduces an overview of the proposed approach. Section 4 describes a household census linking process. The experimental results report in Section 5, and conclude this paper and point out of future directions in Section 6.

2. Related Work

The problems of linking historical census population data came from various parts. These include lack of data quality, huge amount of similar values in full names, address, occupation and ages. It has a more important fact that the situation of residents in a household may change significantly between the decades like birth, death, marriage, moved home, change occupation or change their full name. Consequently, linking households results are not reliable and generated many false matches. It is also a common problem for linking records applications.

In recent years, the modern record linkage methods, which can be applied to meet the problems for historical population census data linking, have been developed by computer science researchers. The probabilistic data cleaning techniques for full names and address which perform than traditional rules-based approaches have been proposed by Christen [7, 4]. An overview of both pattern matching and phonetically encoding based on name matching techniques has been presented.

Zhichun Fu [5] introduced an approach for automatic cleaning and linking of historical census data. This method used household information to link both residents and households across several historical census datasets. The proposed approach has been applied using six census datasets from the United Kingdom between 1851 and 1901.

P. Christen [2] proposed a supervised learning and group linking method to link households with historical census data across time. Firstly, this method figures the similarity between pair of record pairs and uses these results to Supper Vector Machine (SVM) classifier as an input. And then, the SVM classifier classifies the record pairs to a matched and unmatched record pair. They used group linking technique to generate household linking similarities.

It is essential to examine area driven methods for enhancing the historical census record linkage quality. The realizing of the areas social sciences' needs and combines that knowledge with data cleaning and household record linkage methods by the computer science community [7][11].

A group linking method has been applied to generate a household match score by combining similarity scores from matched individual in a household [12]. A Graph matching method [1] was introduced to link households, which takes the structural relationship between household members into consideration.

One problem with the above methods for historical census matching is that matching is performed on the majority of members in a household over a period of time. However, a household may split multiple households due to marriage or movement of another house or may change household structure as birth and death between two censuses. So, the previous proposed methods cannot get accurate household matching results.

3. Overview of Proposed Approach

The proposed approach constitutes two phases as illustrated in Figure. 1. They are record similarity and household graph similarity.

There are three processes in record similarity phase. The first process is attribute similarity calculation by using the approximate string comparison methods. Then, record-pairs similarity is calculated by summing all attributes-wise similarity results. And then, the matched record-pairs are defined from the record-pairs similarity results using the appropriate similarity threshold value [3].

The purpose of the household graph similarity stage is to compute similarities between two graphs. In the construction of household graphs, matched records are used to construct a graph for each household. The graph similarity calculation is then performed based on vertex similarity and edge similarity calculation.

4. Household Census Linking Process

4.1 Attribute Similarity

The historical census datasets contain attributes for each individual in a specific district as detailed in Section 6.

When comparing the records, appropriate approximate string comparison functions have been chosen for each attribute. Before comparing the records, a blocking technique [6] was first applied to reduce the complexity of pair wise linking.

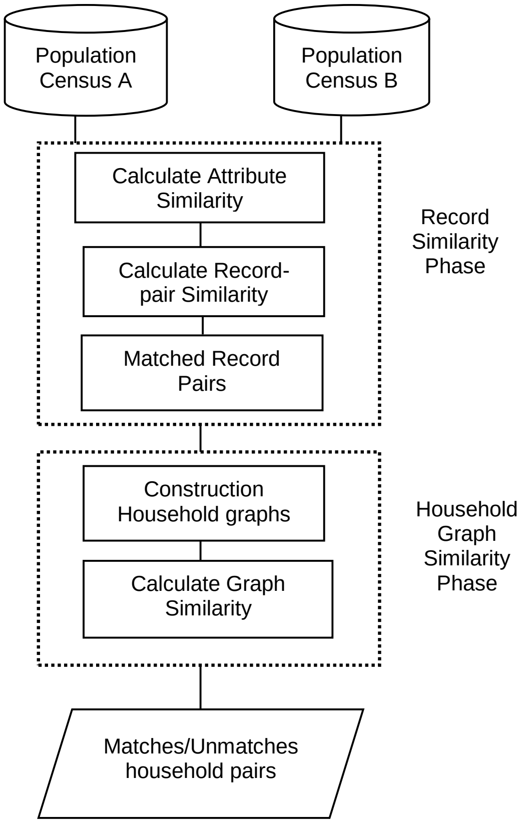

Population Census A
Population Census B
Calculate Attribute Similarity
Calculate Record-pair Similarity
Matched Record Pairs
Record Similarity Phase
Construction Household graphs
Calculate Graph Similarity
Household Graph Similarity Phase
Matches/Unmatches household pairs

Figure 1. Steps of the Proposed Household Graph Matching

The list of attributes and functions used to compute the similarities between values is shown in Table 1.The range of attribute-wise similarities from the records is between 0 and 1. If the score of records is higher, the two attributes are more similar (scores of 1 indicate an exact match, 0 means no similarity).

Table 1. Similarity Method Used for the Five Attributes

Attribute	Method
Surname	Q-gram
First name	Q-gram
Sex	String extract match
Age	Gaussian probability
Address	Longest common subsequence

4.2 Record-pair Similarity

The outputs of the above step are attribute-wise similarities of the selected attributes from the records. The total similarity score Rsim (a, b) was calculated by summing over all attribute-wise similarity scores. The values of Rsim (a, b), total similarity score, are in the range 0 to 5. The higher the total similarity value, the more similar two records are.

We need to determine which record pairs may be true match. We find match record pairs by comparing the total similarity with a threshold ρ , such that

$$\text{Rsim (a, b)} \geq \rho \qquad (1)$$

The linking census data based on the similarity threshold method [3] studied the best appropriate threshold among the five threshold values (2.5, 3.0, 3.5, 4 and 4.5). In this work, threshold values 4 and 4.5 generate only single match record pairs. Threshold value 3.0 generates many false matches. By analyzing the results, threshold value 3.5 covers not only single match record but also multiple match records.

Therefore, we set an appropriate threshold value ρ = 3.5 in our work. After eliminating using threshold value ρ , the small similarity record pairs are moved from the consideration. So, the record pair with the highest similarity can be selected. In some instances, more than one record pairs may have the same highest similarity values, then all of that matched records are selected.

4.3 Household Graphs Construction and Vertex Matching

After record pair selection step, a graph can be constructed for each household. The record matching step can remove a large number of low probability links, such that individual links in a household without high probability do not need to be included in the graph construction. So, this allows small household graphs to be constructed that lead to high computational efficiency.

Figure.2 illustrates an example of the household structure of $H_{1851}$from 1851 Census. Figure.3 also shows the structural information of two households (H_{1861}- A and H_{1861}- B) from 1861 Census. The individuals are associated to a single household in each dataset. A household (H_{1851}) in 1851 splits two households (H_{1861}-A and H_{1861}-B) in 1861 due to marriage.

When constructing household graphs, vertices are corresponding to the household members and edges are connecting between vertex pairs. The proposed approach considered three edges attributes: age difference, generation difference and role-pair between individuals in the household. For instance, as shown in Figure 1, a record with role value “wife” is in the same generation with the "head of household", so their generational difference is 0. The value of generational difference between "head of household" and "son" or "daughter" is 1. The age difference is the difference age values between head of family and household members in a household. For example, edge value of "27" in Figure 2 is the difference age values of head (r11) and his member (r14).

H_{1851}	SUR NAME	FIRST NAME	Relationship to Head	SEX	AGE	STREET	COUNTRY
r11	rickard	Thomas	head	M	40	Aughnacur	Cavan
r12	rickard	Kate	wife	F	38	Aughnacur	Cavan
r13	rickard	William	son	M	18	Aughnacur	Cavan
r14	rickard	James	son	M	13	Aughnacur	Cavan
r15	rickard	kathleen	daughter	F	20	Aughnacur	Cavan
r16	rickard	Peter	father	F	60	Aughnacur	Cavan
r17	rickard	Bridget	mother	M	59	Aughnacur	Cavan

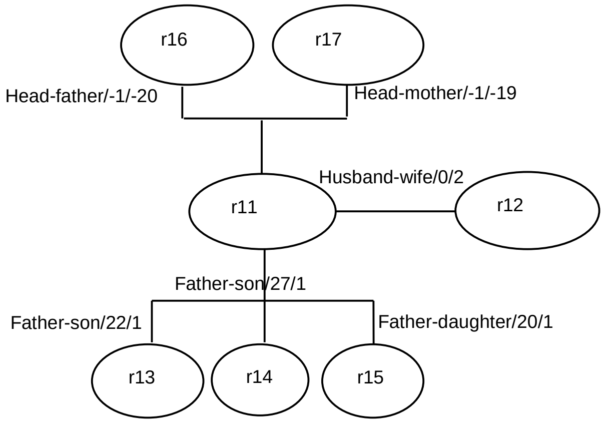


Figure 2. An Example of a household (H_{1851}) from 1851 census

Several target records may be included in the record matching step. Therefore, one-to-many and many-to-one vertex matching may be generated between two graphs. Then, the optimal vertex to vertex has to be determined. The vertex matching was calculated by maximizing the sum of matched records probabilities.

4.4 Graph Similarity

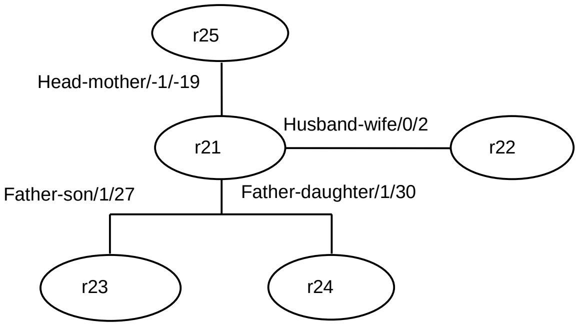
In the record linking step, a record may be linked to several records in different households. Therefore, a graph containing the record may be linked to several other graphs. Similar to the record matching step, decisions also have to be made on which graph pair is a possibly a true match, and if there are multiple matches, which pair is the correct one. So, this requires the calculation of graph similarity. We define the similarity between G and G′ as

$$f(G,G') = f(V,V') + f(E,E') \quad (2)$$

where $f(V,V')$ and $f(E,E')$ are the total vertex similarity and total edge similarity.

The vertex similarity has been generated in the record similarity step. Let $sim_v(r_i, r_i')$ be the vertex similarity of the ith record pair in the graph, and the total number of vertices in G be N, then

H_{1861}-A	SUR NAME	FIRST NAME	Relationship to Head	SEX	AGE	STREET	COUNTRY
r21	rickard	thomas	head	M	50	Aughnacur	Cavan
r22	rickard	kate	wife	F	48	Aughnacur	Cavan
r23	rickard	james	son	M	23	Aughnacur	Cavan
r24	rickard	kathleen	daughter	F	20	Aughnacur	Cavan
r25	rickard	bridget	mother	F	69	Aughnacur	Cavan



H_{1861}-B	SUR NAME	FIRST NAME	Relationship to Head	SEX	AGE	STREET	COUNTRY
r31	Rickard	William	Head	M	28	Aughnacur	Cavan
r32	Reilly	Ellens	Wife	F	20	Aughnacur	Cavan
r33	Rickard	Lusei	daughter	M	3	Aughnacur	Cavan

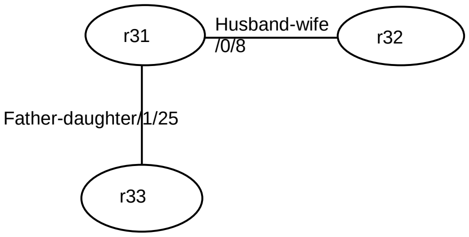


Figure 3. An Example of two households (H_{1861}-A and H_{1861}-B) from 1861 census

$$f(V,V') = \frac{\sum_{i=1}^{N} sim(r_i, r_{i'})}{N} \quad (3)$$

Let r_{ijk} be the kth ($k \epsilon [1, \ldots K]$) attribute of the edge e_{ij} which connects record i and record j in graph G. The edge similarity calculation is defined as

$$sim(r_{ij}, r'_{ij}) = \frac{\sum_{k=1}^{K} sim_a(r_{ijk}, r\prime_{ijk})}{K} \quad (4)$$

The total edge similarity calculation is based on differences on edge attributes between each pair of edges in the graph pair.

$$f(E,E')=\frac{\sum_{i=1}^{L} sim(r_{ij},r'_{ij})}{L} \qquad (5)$$

where L is the number of edges in the household graph.

The calculation of graph similarity allows determining the optimal match from several household graphs.

$$f(G,G') > \alpha \qquad (6)$$

If the graph similarity is larger than threshold value α then it is examined as true match. The parameter α learned from the training dataset.

5. Experimental Result

This section provides the evaluation of the proposed graph-based approach. Two Ireland historical census datasets [15] are used, which are collected from the district of Aghullaghy in Cavan in Ireland for the period of 1901 and 1911.

There are twelve attributes for each record, first name, surname, age, sex, relation to head, religion, birth place, occupation, literacy, Irish language, marital status and specific illnesses. These data were standardized and cleaned before applying the record and household linkage process [5].

The proposed method (Graph Similarity) was compared to other baseline methods (Highest Similarity and Vertex Similarity). Highest Similarity, the first baseline, the method calculates household similarity based on the highest similarity scores. If one household is linked to several target households in another dataset with the highest record similarity score is selected.

Based on the linked records, household graphs were built in Vertex Similarity, the second baseline, method. Then, household matching is determined only by the vertex similarity calculation in Equation (3). This is equal to the calculating of the suitable record similarity on those records to build household graph.

Table 2 shows the total household pairs and the number of matched household pairs with different similarity methods. Highest Similarity generates 22 matched households of 265 household pairs. It matches only a household in one dataset to one household in another dataset. Vertex similarity causes 58 pairs of total household pairs. It provides multiple matches of a household in another dataset. However, it includes false multiple household matches.

The proposed method, Graph Similarity, generates 38 matched pairs of total 265 pairs, considers the relationships between members in a household. Therefore, it covers single matched and multiple matched household pairs.

Table 2.Total Household pairs with different Similarity Methods

Similarity Methods	Total household pairs	Number of matched household pairs
Highest Similarity	265	22
Vertex Similarity	265	58
Graph Similarity	265	38

Table 3.Comparison of performance of the proposed method and other baseline methods

Similarity Methods	Precision	Recall	F-score
Highest Similarity	0	0.415	0.587
Vertex Similarity	0.862	0.926	0.893
Graph Similarity	0.974	0.974	**0.974**

The precision, recall and F-score were calculated for similarity methods. The results from the similarity methods being compared are summarized in Table 3.It shows that the graph similarity method has generated the best F-score among the other similarity methods. Figure 3 shows the performance comparison for household linking.

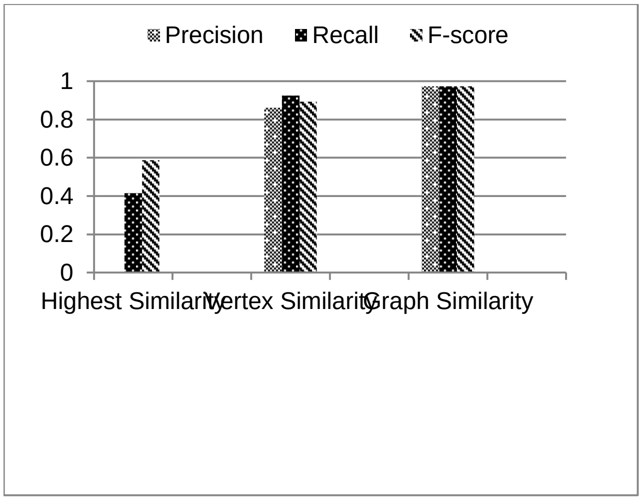


Figure 4. Performance Comparison for household linking

This figure presents the graph similarity methods outperformed than highest similarity and vertex similarity.

The results that the proposed method is effective reduce number of incorrect links and support multiple household linking between two years interval.

6. Conclusion

This paper has been introduced a graph matching approach to match households for population census data. The aim is to decrease ambiguous links and match multiple households over a certain period of time. This approach considers not only record similarity but also incorporates the relationships into the household matching step. The household graph linking process is executed in two phases. The first phase computes pair-wise record linking based on the total attributes similarity values. After record pairs similarities are computed, matches or un-matches are classified by setting appropriate threshold values. The second phase is household graph matching. Household graphs are constructed by using the matched record pairs. The experimental results have shown that the relationship between individuals in a household is very useful in household matching. The proposed method can generate very reliable linking outcomes for both single and multiple household linkages.

We will study graph matching learning method on large dataset and incorporate more features for graph similarity method.

7. References

[1] Z. Fu, P. Christen, and J Zhou, "A Graph Matching Method for Historical Census Household Linkage", in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2014, pp. 485-496.

[2] Z. Fu, P. Christen and M. Boot, "A Supervised Learning and Group Linking Method for Historical Census Household Linkage," in *AusDM'11 Proceedings of the Ninth Australasian Data Mining Conference*, Ballarat, Australia, vol. 121, pp. 153-162, December 01 – 01 2011.

[3] Khin Su Mon Myint, Thet Thet Zin and Kyaw May Oo, "Analysis of Historical Census Household data with Similarity Threshold", *ICAIT (The 1st International Conference on Advanced Information Technologies*, Myanmar, 2017.

[4] Z. Fu, P. Christen and M. Boot, "Automatic Cleaning and Linking of Historical Census Data using Household Information," in *11th IEEE ICDM (International Conference on Data Mining) Workshop*, 2011, pp. 413–420.

[5] Z. Fu, H.M. Boot, P. Christen and J. Zhou, "Automatic Record Linkage of Individuals and Households in Historical Census Data," in *International Journal of Humanities and Arts Computing*, vol. 8, no. 2, pp. 204-225, 2014.

[6] P. Christen, "Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection", *Springer*, 2012.

[7] P. Christen, "Development and user experiences of an open source data cleaning, de duplication and record linkage system," in *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 39-48, 2009.

[8] Dmitri. V. Kalashnikov and S. Mehrotra, "Domain-independent data cleaning via analysis of entity-relationship graph," *ACM Transactions on Database Systems Journal*, vol. 31, no. 2, pp. 716-767, June 2006.

[9] B. - W. On, N. Koudas, D. Lee, and D. Srivastava, "Group linkage," in *Proceedings of the IEEE 23rd International Conference on Data Engineering*, 2007.

[10] E. Fure, "Interactive record linkage: The cumulative construction of life courses Demographic Research," vol. 3, no. 11, December 2000.

[11] S. Ruggles, "Linking historical censuses: a new approach," *History and Computing*, vol. 14, no. 1+2, pp. 213–224, 2006.

[12] G. Bloothooft, "Multi-source family reconstruction," *History and Computing* ,vol. 7, no. 2, pp. 90–103, 1995.

[13] D. Quass and P. Starkey, "Record linkage for genealogical databases," in *ACM SIGKDD 2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, Washington DC, August 24-27, 2003.

[14] A. Ashkpour, K. Mandemeakers and A. Meronopenuela, "The Aggregate Dutch Historical Census Historical Methods," vol. 48, no. 4, October 2015.

[15] http://www.census.nationalarchives.ie/

Building Ontology for Big Data in Column-oriented NoSQL Database

Nang Kham Soe, Tin Tin Yee, Ei Chaw Htoon
University of Information Technology
Yangon, Myanmar
nangkhamsoe@uit.edu.mm, tintinyee@uit.edu.mm, eichawhtoon@uit.edu.mm

Abstract

Big data are usually integrated data from different sources in a structured, semi-structured and unstructured data collection. NoSQL databases are in the base of big data storage. These databases suffer from lack of semantics since they are able to handle unstructured data. In order to solve that issue, an ontology-based representation of the information stored in NoSQL databases is highly needed. This paper proposes an approach to construct an ontological form of big data stored in column-oriented NoSQL database namely Cassandra. The approach defines mapping rules and builds ontology by applying these rules.

Keywords- Ontology, Semantic Web, Big data, NoSQL, Cassandra

1. Introduction

Big data was defined as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” Big data often ranges from a few dozen terabytes (TB: approximately 1012 bytes) to multiple petabytes (PB: approximately 1015 bytes) [5]. Big data is often represented by large amounts of high-dimensional and poorly structured or organized forms when the data is typically generated from heterogeneous sources. It can be either structured (e.g. spreadsheets, relational databases), unstructured (e.g. text, image), and/or semi-structured such as radio frequency identification (RFID) data and extensible markup language (XML) data [9].

NoSQL databases are in the base of big data storage. They permit to store large volumes of structured, semi-structured, and unstructured data. These databases suffer from lack of semantics since they are able to handle unstructured data. Cassandra is a type of column-oriented NoSQL database. It has no fixed schema. In this paper, an ontology based approach is proposed to extract hidden semantics from schema-less data store. The proposed approach defines mapping rules to represent a data source as an ontological form. The proposed system uses OWL (Web Ontology Language) to describe ontology as the target schema and data source corresponds to column-oriented database namely Cassandra. The system consists of three steps: understanding data source model, defining mapping rules, and building ontology.

The rest of the paper is organized as follows. Section 2 introduces about Cassandra database. Section 3 describes about ontology. Related works are presented in Section 4, and we explain our proposed system in Section 5. Finally, we draw conclusions in Section 6.

2. Column-oriented NoSQL DB: Cassandra

The concept of non-relational databases (NoSQL DBs) refers to a database alternative to the relational model that arranges data discretely into tables of columns and rows. NoSQL DBs are commonly associated with more flexible deployment, high read/write performance as well as scaling to very large data sets. There can be distinguished four different kinds of NoSQL DBs [10]: document databases (examples: Couchbase, MongoDB), key-value stores (examples: Riak, Amazon’s Dynamo), graph databases (examples: InfoGrid, Infinite Graph, Neo4J), and column-oriented stores (examples: HBase, Cassandra).

Apache Cassandra is the type of column oriented NoSQL database. It is a free and open-source , a highly scalable, high-performance distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.

The data model of Cassandra is wide column store, and, as such, essentially a hybrid between a key-value and a tabular database management system. Rows are organized into tables; the first component of a table's primary key is the partition key; within a partition, rows are clustered by the remaining columns of the key.

A column family or a super column family (called "table" since CQL 3) resembles a table in an RDBMS. Column families contain rows and columns. A super column family consists of a row key and a number of super columns. Each row is uniquely identified by a row key. Each row has multiple columns, each of which has a name, value, and a timestamp. Unlike a table in an RDBMS, different rows in the same column family do not have to share the same set of columns, and a column may be added to one or multiple rows at any time. In

Cassandra, it also has special column called super column that stored a map of sub-columns. The data model of Cassandra is as shown in Figure 1.

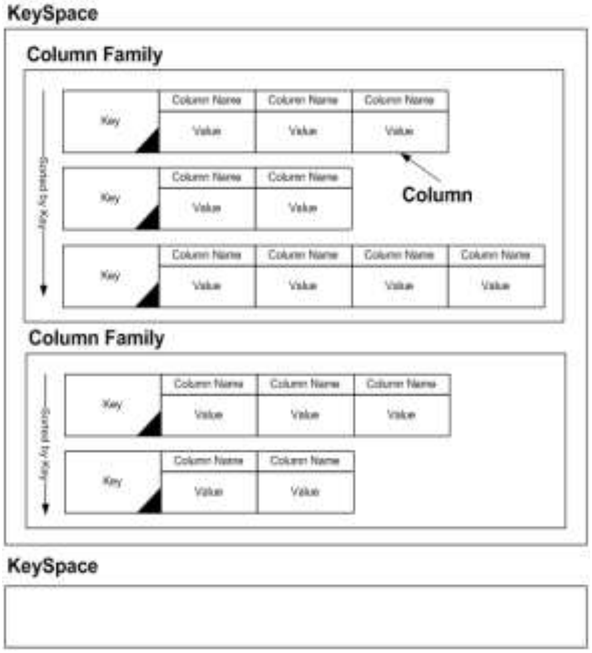


Figure 1. Cassandra data model

3. Ontology

Ontology [6] is the core of the Semantic Web technology, which is also an important method to conceptualize domain knowledge and modeling that can be used to describe the semantics of the data. The potential role of semantics in big data [7] is to describe data source ’schemas using ontologies. Ontologies, as formal models of representation with explicitly defined concepts and named relationships linking them. To make data schematically comparable, for each one of data sources would be represented by its local ontology. In NoSQL databases, parts of structure definition are mixed with the data (e.g..Keys in key-value format) and some are unstructuredness. The lack of formal schema in NoSQL databases can be compensated by ontologies. There are different languages to describe the ontology. Different ontology languages provide different facilities. OWL is the World Wide Web Consortium (W3C) recommended standard ontology description language.

There are three main components in OWL [4]: Classes, Properties and Individuals. Classes are a concrete representation of concepts. Classes may be organized into a superclass-subclass hierarchy, which is also known as taxonomy. OWL Properties represent relationships. There are two main types of properties, Object properties and Datatype properties. Object properties are relationships between two individuals. Datatype properties describe relationships between individuals and data values. Individuals represent objects in the domain and are also known as instances. Individuals can be referred to as being ‘instances of classes’.

4. Related Works

Abbes et. al intended to develop a novel approach for ontology learning from a NOSQL database to manage big data about a specific domain [1]. The paper introduced links between ontologies and Big Data. They presented transformation rules for building OWL ontologies from NOSQL database. They transform all possible cases in the MongoDB database into ontological constructs. It is divided into four main steps to learn ontology from MongoDB. The first step is the creation of the ontology skeleton. It consists of class definition and subsumption relation detection between classes. The second step is to learn objectProperties and dataTypeProperties. In the third step, individuals are identified. Finally, in the fourth step, it deduces axioms and constraints.

Abbes et. al proposed ontology based big data integration approach based on a NoSQL database, namely MongoDB and modular ontologies. It follows three steps: wrapping data sources to MongoDB databases, generating local ontology corresponding to each data source, and composing the local ontologies to get a global one. The target schema is represented as OWL ontology [2].

In [8], the authors proposed a semantic Extract-Transform-Load (ETL) framework for big data integration. The proposed framework uses semantic technologies to integrate and publish data from multiple sources as open linked data. The use of semantic technologies is introduced in the Transform phase of an ETL process to create a semantic data model and generate semantically linked data (RDF triples) to be stored in a data mart or a data warehouse.

Kiran et. al presented ontology based semantic integration system for a column-oriented NoSQL data store like HBase [3]. It follows three steps: schema extraction for each data source based on two methods: On-Line schema generation and Offline schema generation, schema to ontology conversion and ontology alignment-merging. The proposed system represents RDF triples as the target schema.

5. Proposed System

In big data environment, data come from different data sources. NoSQL databases are used as big data stores. These databases suffer from lack of semantics. In order to solve the issue, ontology based approach to extract hidden semantics from big data is proposed. The objective of the proposed system is to later construct an ontological form (global ontology) for big data from different data stores. As the first step, this paper mainly focuses on building local ontologies for each of data stores. The proposed system deals with Cassandra data stores as input. The system consists of three steps: understanding the data source model, defining mapping rules, and building ontology as shown in Figure 2.

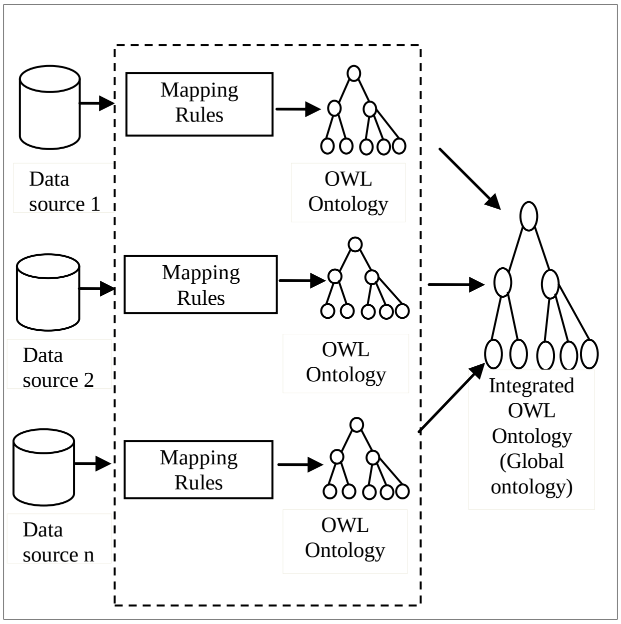


Figure 2. Proposed system architecture

5.1. Understanding Data Source Model

The proposed system uses NoSQL database, namely Cassandra. As mentioned in Section 2.1, outermost container for data in Cassandra is keyspace. Each keyspace has at least one and often many column families. A column family is a container of a collection of rows. Each row contains ordered columns. In Cassandra, it has special columns called super columns that stored a map of sub-columns.

5.2. Defining Mapping Rules

To generate semantics from Cassandra database, it is mapped to the OWL ontology by means of the following rules.

Rule 1: Defining Class

Each column family in the keyspace is transformed into OWL class.

Rule 2: Defining Datatype Property

The value of column may have different data types, namely basic data types (int, string, etc.), collection types (map, set, list) and user-defined data types. If the column has basic data types, that column is transformed into a dataTypeProperty for which "domain" the class corresponding to the column family that contains that column.

Rule 3: Defining Object Property

If the column has collection types or user-defined data types, it is transformed into an object property for which "domain" the class corresponding to the column family that contains the column and "range" is the class corresponding to the collection or user-defined type.

Rule 4: Defining Individuals

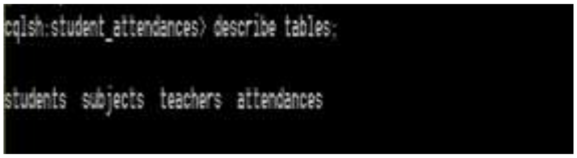
The value of each column in the row is transformed into individuals in the ontology.

Rule 5: Defining Axioms

If two different rows in the column family contain the same set of data values for all columns, it deduces two different classes corresponding to these two different rows are equivalent. Otherwise, it deduces these two different rows are disjoint.

5.3. Building OWL Ontology

As a test dataset, students' attendance data of a university is used to show ontology conversion from data stored in Cassandra. The dataset consists of four column families (tables) as depicted in Figure 3.



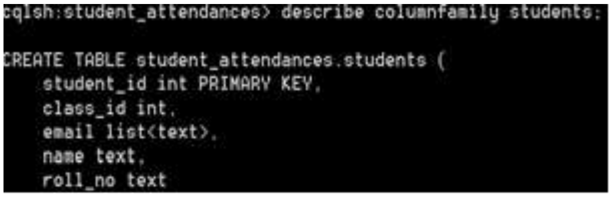
```
cqlsh:student_attendances> describe tables;

students  subjects  teachers  attendances
```

Figure 3. Test cassandra dataset

According to Rule 1, all column families are transformed to ontology class.

<owl:Class rdf:ID="attendances"/>
<owl:Class rdf:ID="students"/>
<owl:Class rdf:ID="teachers"/>
<owl:Class rdf:ID="subjects"/>

According to Rule 2, the column that has basic data types is transformed into a dataTypeProperty. The following only gives dataType property transformation for the column family "students" due to space limitation. The internal structure of "students" is as shown in Figure 4. Three columns have basic data type and the column "email" has collection datatype.



```
cqlsh:student_attendances> describe columnfamily students;

CREATE TABLE student_attendances.students (
    student_id int PRIMARY KEY,
    class_id int,
    email list<text>,
    name text,
    roll_no text
```

Figure 4. Column family ”students”

```
<owl:DatatypeProperty rdf:ID="roll_no">
<rdfs:range rdf:resource="http://www.w3.org/2001/
XMLSchema # string"/>
<rdfs:domain rdf:resource="#students"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="class_id">
 <rdfs:range rdf:resource="http://www.w3.org/2001/
XMLSchema#integer"/>
<rdfs:domain rdf:resource="#students"/>
</owl:DatatypeProperty>
 <owl:DatatypeProperty rdf:ID="name">
 <rdfs:range rdf:resource="http://www.w3.org/2001/
XMLSchema#string"/>
<rdfs:domain rdf:resource="#students"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="class_id">
<rdfs:range rdf:resource="http://www.w3.org/2001/
XMLSchema#integer"/>
<rdfs:domain rdf:resource="#students"/>
</owl:DatatypeProperty>
```

According to the Rule 3, the column “email” in the column family “students” is transformed into object property.

```
<owl:ObjectProperty rdf:ID="hasemail">
   <rdfs:domain rdf:resource="#students"/>
   <rdfs:range rdf:resource="#email"/>
 </owl:ObjectProperty>
```

The individual is an instance of the class in the ontology. Accordance with Rule 5, the value of each column is transformed to individual in the ontology. The following only gives individual transformation for one record of the column family “students” due to space limitation.

```
<students:student_id  rdf:about="#82">
    < students:roll_no>5SE-1 </ students:roll_no>
    < students:class_id>4</ students: class_id >
    < students:name>Juzi Hein </ students:name>
    < students:email>juzihein@uit.edu.mm
    </ students: email >
  </ students:students_id>
```

6. Conclusion

In big data environment, data come from different data sources. NOSQL databases are in the base of storage of big data. These databases suffer from lack of semantics since they are able to handle unstructured data. Therefore, an ontology-based representation of the information stored in NOSQL databases is highly needed. The paper intends to develop an approach for building ontology from big data. The proposed system mainly focuses on hidden semantics extraction from data stored in column-oriented database namely Cassandra.

The next step of our research tries to build ontologies for different data model and focuses on semantic heterogeneity problem in the big data integration process.

7. References

[1] Abbes, H., Boukettaya, S., Gargouri, F, “Learning ontology from Big Data through MongoDB database.” , Proceedings of IEEE/ACS 12th International Conference of Computer Systems and Applications, 2015, pp. 1–7.

[2]Abbes, H., Gargouri, .F, “Big Data Integration: a MongoDB Database and Modular Ontologies based Approach”, 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2016), 5-7 September 2016.

[3] Kiran V K, Vijayakumar R, “Ontology Based Data Integration of NoSQL Datastores”, 9th International Conference on Industrial and Information Systems (ICIIS), 15-17 Dec. 2014

[4] Matthew Horridge, Simon Jupp, Georgina Moulton, Alan Rector, Robert Stevens, Chris Wroe, “A Practical Guide To Building OWL Ontologies Using Prot´eg´e 4 and CO-ODE Tools Edition 1.1”

[5] Minelli, M., Chambers, M. & Dhiraj, A., “Big data, big analytics: emerging business intelligence and analytic trends for today's businesses.” , 2012.

[6] Li Kang, Li Yi, LIU Dong, “ Research on Construction Methods of Big Data Semantic Model”, Proceedings of the World Congress on Engineering 2014 Vol I,WCE 2014, July 2 - 4, 2014.

[7] Marina P., Boris V., “Semantic Web Technologies and Big Data Warehousing”, MIPRO, May 21-25, 2018.

[8] Srividya K Bansal, “Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration,” IEEE International Congress on Big Data, 2014.

[9] Wang Lidong, Randy Jones, “Big Data Analytics for Disparate Data”, American Journal of Intelligent Systems, 7(2): 39-46 (2017).

[10] Woodey A., “Forrester Ranks the NoSQL Database Vendors. Datanami”, 2014. Available: http://www.datanami.com/2014/10/03/forrester-ranks-nosql-database-vendors/.

Evaluation of Apache Kafka in Real-Time Big Data Pipeline Architecture

Thandar Aung, Hla Yin Min, Aung Htein Maw
University of Information Technology
Yangon, Myanmar
thandaraung@uit.edu.mm, hlayinmin@uit.edu.mm, ahmaw@uit.edu.mm

Abstract

Today, many applications based on real-time analytics need to enable time-critical decision with real-time requirements and process with reliability requirements. Modern distributed systems are growing exponentially as far as performance and scale. The main purpose of Big Data real-time processing is to realize an entire system that can process such mesh data in a short time. And the performance of processing time can be guaranteed in a satisfactory range. To develop a distributed data pipeline, the system proposes real-time big data pipeline by using Apache Kafka and Apache Storm. Apache Kafka is currently the most popular framework used to ingest the data streams into the processing platforms. However, there are many challenges how to send reliable messages on many servers. This paper focuses on the comparison of the processing time between successful processes and failed processes on many servers. The experimental results show the performance impact on both producer and consumer of the Apache Kafka framework.

Keywords- Apache Kafka, Apache storm, Asynchronous replication, Real time processing,

1. Introduction

In the present big data era, the very first challenge is to collect the data as it is a huge amount of data and the second challenge is to analyze it. This analysis typically includes User behavior data, Application performance tracking, Activity data in the form of logs and Event messages. Processing or analyzing the huge amount of data is a challenging task. It requires a new infrastructure and a new way of thinking about the business and IT industry works. Today, organizations have a huge amount of data and at the same time, they have the need to derive value from it.

Real-time processing involves continual input, process and output, and minimal response time. Real-time processing is used when acting within a very short period of time is significant, so this type of processing allows the organization or system to act immediately. Fast response time is critical in these examples: bank ATM operations, money transfer systems, aircraft control or security systems. Real-time processing is fast and prompt data processing technology that combines data capture, data processing and data exportation together. Real-time analytics is an iterative process involving multiple tools and systems. It consists of dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use. Real-time information is continuously getting generated by applications (business, social, or any other type), and this information needs easy ways to be reliable and quickly routed to multiple types of receivers. This leads to the redevelopment of information producers or consumers to provide an integration point between them. Therefore, a mechanism is required for seamless integration of information of producers and consumers to avoid any kind of rewriting of an application at each end. Real-time usage of these multiple sets of data collected from production systems has become a challenge because of the volume of data collected and processes. Kafka has high throughput, built-in partitioning, replication, and fault tolerance, which makes it a good solution for large-scale message processing applications.

This paper has proposed a real-time processing pipeline using the open-source frameworks that can capture a large amount of data from various data sources, process, store, and analyze the large-scale data efficiently. This proposed system evaluates the performance impact on Apache Kafka to develop big data pipeline architecture by using Apache Kafka and Apache Storm. We performed several experiments to prowle the performance testing of the system under different servers.

The remainder of this paper is organized as follows: Section 2 reviews the related work of this paper. Section 3presents the system architecture to develop a real-time messaging system. Section 4 describes the experimental setup of the system and experimental results for the large messaging system in big data pipeline architecture. Section 5 describes the conclusion and future work.

2. Related Work

Seema Balhara, Kavita Khanna [1] has shown that Kafka has superior performance when compared to two popular messaging systems. The author uses ActiveMQ and RabbitMQ. The author describes that Kafka achieves much higher throughput than the conventional messaging system.

Hassan Nazeer, Waheed Iqbal, Fawaz Bokhari[2] Faisal Bukhari, Shuja Ur Rehman Baig has proposed to evaluate their proposed real-time text processing pipeline using open-source big data tools. This paper intends to minimize the latency to process data streams and conduct several experiments to determine the performance and scalability of the system against different cluster size and workload generator.

Martin Kleppmann [3] explains the reasoning behind the design of Kafka and Samza, which allow complex applications to be built by composing a small number of simple primitives – replicated logs and stream operators. These draw parallels between the design of Kafka and Samza, batch processing pipelines, database architecture and design philosophy of UNIX.

Muhammad Syafrudin [8]explains capable of processing a massive sensor data efficiently when the number of sensor data and devices increases. This paper is expected to support the management in their decision-making for product quality inspection and support manufacturing sustainability. The OSRDP used several open source-based big data processing such as Apache Kafka, Apache Storm and MongoDB.The results showed that the proposed system is capable of processing a massive sensor data effeciently when the number of sensors data and devices increases.

Paul Le Noac'h, Alexandru Costan, Luc Boug´e[4]has proposed the evaluation of several configurations and performance metrics of Kafka and studied how the ingestion performance can impact the overall stream processing.

Wenjie Yang, Xingang Liu and Lan Zhang [5] have also proposed to ensure the practical applicability and high efficiency, to show acceptable performance in the simulation. The paper showed an entire system RabbitMQ, NoSQL and JSP are proposed based on Storm, which is a novel distributed real-time computing system. The paper organized a big data real-time processing system based on Strom and other tools.

Mohit Maske, Dr. Prakash Prasad, International Journal of Advanced [6] intends to ensure the practical and high efficiency in the simulation system that is established and shown acceptable performance in various expressions using data sheet. It proved that data analysis system for stream and real-time processing based on storm can be used in the various computing environments.

Overall, Kafka has weak guarantees as a distributed messaging system. There's no ordering guarantee when the messages are coming from different partitions. The weak point of related papers is to develop the performance of processing in the pipeline. This paper has thoroughly evaluated the performance of producer and consumer in Kafka. The results show that higher performance can lead to significant performance improvements of the real-time messaging system.

3. System Architecture for Real Time Messaging System

This section focuses the performance of producer and consumer in Apache Kafka processing on the pipeline architecture. The performance of Kafka processing modifies to be more reliable on the pipeline architecture in Figure 1.

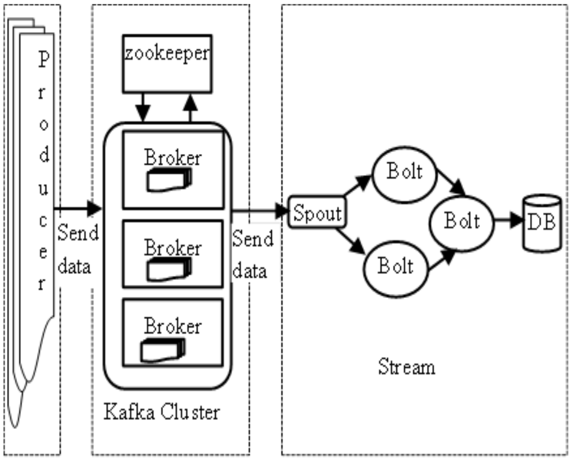


Figure 1. Real-time big data pipeline Architecture

A producer publishes messages to a Kafka topic by using asynchronous type. Brokers can divide messages into many partitions. Zookeeper serves as the coordination interface between the Kafka broker in topic and consumers. The Kafka spout uses the same Zookeeper instance that is used by Apache Storm, to store the states of the message offset and segment consumption tracking if it is consumed. Kafka can act as a buffer or feeder for messages that need to be processed by Storm. Kafka and Storm naturally complement each other, and their powerful cooperation enables real-time streaming analytics for fast-moving big data. The paper intends to evaluate the performance of Apache Kafka in the pipeline architecture.

3.1 Apache Kafka Architecture

Kafka [9] is an open source, distributed publish-subscribe messaging system. A producer publishes messages to a Kafka topic that is created on a Kafka broker acting as a Kafka server. Kafka maintains feeds of messages in categories called topics. Kafka is run as a cluster comprised of one or more servers, each of which is called a broker. Brokers can divide messages into many partitions. Each partition is optionally replicated across a configurable number of servers for fault tolerance. Producers send messages over the network to the Kafka cluster which in turn serves them up to consumers. Brokers and consumers use Zookeeper to get the state information and to track messages offsets, respectively.

Kafka supports two replication modes in processing: Synchronous replication: A message to be published is acknowledged as soon as it reaches 1 replica. Asynchronous replication: The only difference in this mode is that, as soon as a lead replica writes the message to its local log, a message is only acknowledged after it reaches multiple replicas. But, as a downside, this mode does not ensure message delivery in case of broker failure. In a Kafka cluster, each server plays a dual role; it acts as a leader for some of its partitions and also a follower for other partitions. The leader is responsible for handling all read and write requests for the partition while the followers asynchronously replicate data from the leader.

If any of the follower in-sync replicas fail, the leader drops the failed follower from its ISR (in-sync replicas) list. After the configured timeout period will continue on the remaining replicas in ISR (in-sync replicas) list. As soon as the follower becomes fully synced with the leader, the leader adds it back to the current ISR list. If the leader fails, the process of choosing the new lead replica involves all the followers' ISRs registering them with Zookeeper. The very first registered replica becomes the new lead replica and its log end offset (LEO) becomes the offset of the last commit. The rest of the registered replicas become the followers of the newly elected leader.

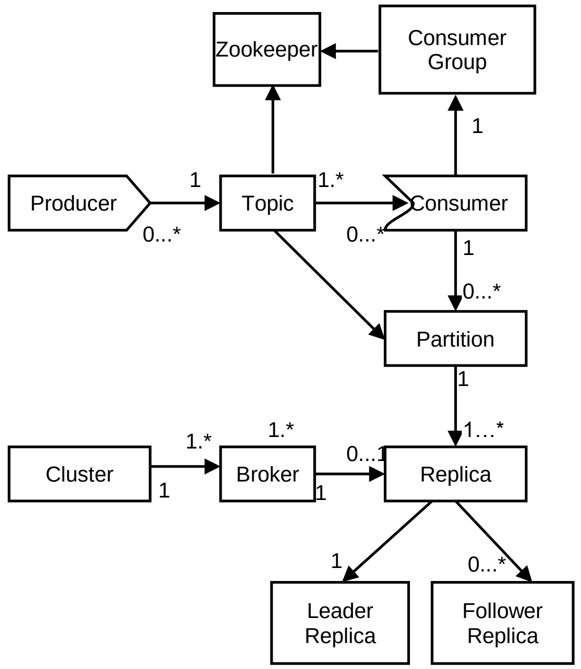


Figure 2. Processes in Kafka

Figure2 describes the processing of an Apache Kafka flow. Firstly, a producer sends a message to one topic at a time. A topic has 0 or more producers. A consumer subscribes to one or more topics. A topic has zero or more consumers. A consumer is a member of one consumer group. Zookeeper serves as the coordination interface between the Kafka broker in topic and consumers. A partition has one consumer per group. A consumer pulls messages from zero or more partitions per topic. A topic is replicated over one or more partitions. A partition has one or more replica. A replica is on one broker. A broker has zero or one replica per partitions. A partition has one leader and zero or more followers.

Zookeeper [12] is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. Zookeeper is also a high-performance coordination service for distributed applications. Specifically, when each broker or consumer starts up, it stores its information in a broker or consumer registry in Zookeeper. The broker registry contains the broker's hostname and port, and the set of topics and the partitions stored on it. The service itself is distributed and highly reliable.

3.2 Apache Storm

Storm [10] is also an open source, distributed, reliable, and fault tolerant system for processing streams of large volumes of data in real-time. A Storm cluster follows a master-slave model where the master and slave processes are coordinated through Zookeeper. The Nimbus node is the master in a Storm cluster. Supervisor nodes are the worker nodes in a Storm cluster .In Storm terminology, [10] a topology can be represented by a direct acyclic graph, where each node does some kind of processing and forwards it to the next node(s) in the flow. The topology of the storm is described as a part of the pipeline architecture as in Figure 1.The followings are the components of a Storm topology:

Stream*:* A stream is an unbounded sequence of tuples that can be processed in parallel by Storm. Each stream can be processed by a single or multiple types of bolts.

Spout*:* A spout is the source of tuples in a Storm topology. The spout is the input stream source which can read from the external data source.

Bolt: The spout passes the data to a component called a bolt. Bolts are processor units which can process any number of streams and produce output streams. A bolt is responsible for transforming a stream.

4. Experimental Setup and Results

The experimental setup is performed by using two open source frameworks Apache Kafka 2.11-0.9.0.0 and Apache Storm 0.9.7 as the main pipeline architecture. JAVA 1.8 is running on underlying pipeline architecture. The Apache Maven 3.5.0uses as in Kafka-Storm integration. The Real-time data for experiments are mobile phone spam messages from the Grumble text Web site.

Table 1.Hardware Specification

Operating system	Windows 32-bit Operating system
RAM	4.00 GB
Hard-disk	1 TB
Processor	Intel(R) Core(TM) i7-4770 CPU @3.40GHz

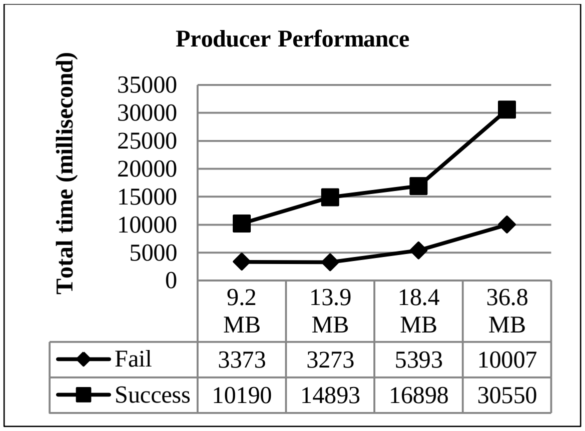
The evaluation of overall Kafka processing in pipeline architecture is as follows:

1. Start Zookeeper server for processing.
2. Start Kafka local server to define broker id, port, and log dir.
3. Create a topic to show a successful creation message.
4. Producer publishes them by asynchronous type as messages to the Kafka cluster.
5. The consumer consumes messages from Zookeeper.
6. Check the leader and followers in ISR list and replica in Zookeeper.
7. Get process id and kill leader in one server.
8. Check total messages in processing by using GetOffsetShell tools
9. Evaluate the performance of producer running alone by using producer performance tools.
10. Evaluate the performance of the consumer by using consumer performance tools.
11. Compare the performance by testing the successful and failed processes
12. Get some messages and run Kafka-Storm integration pipeline.

There are four different experiments that are conducted to evaluate the performance over Kafka processing.

Table 2.Summery of Experiments

Experiment	Description
1.Producer Performance	Deployed Apache Kafka on two servers, 40 partitions, two replication factor, different types of dataset with five batch size
2.Consumer Performance	Deployed Apache Kafka on two servers, 40 partitions, two replication factors and different types of dataset with five batch size
3.Producer Performance	Deployed Apache Kafka on three servers, 40 partitions ,three replication factors, different types of dataset with five batch size
4.Consumer Performance	Deployed Apache Kafka on three servers, 40 partitions ,three replication factor and different types of dataset with five batch size

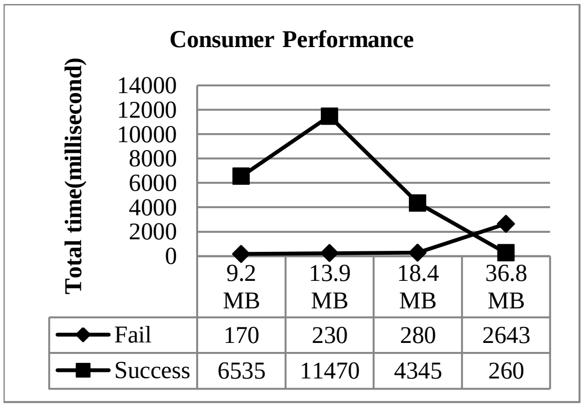


	9.2 MB	13.9 MB	18.4 MB	36.8 MB
◆ Fail	3373	3273	5393	10007
■ Success	10190	14893	16898	30550

Figure 3. Producer Performance on two servers

A. Experiment 1:Producer Performance on two servers

Figure 3 shows a single producer to publish various data sizes of messages 9.2 MB, 13.9 MB, 18.4 MB, 36.8 MB respectively. It configured the Kafka producer to send messages asynchronously by five batches size. Figure 3 shows the result. The x-axis represents the amount of data sent to the broker, and the y-axis corresponds to the total time in producer performance. It compares the total times on the difference between successful and failed processes with 40 partitions in various data sizes on two servers. We consider this experiment as total time by calculating based on producer performance tools in Apache Kafka. The successful processing time increases about three times than the failed process. There are a few reasons why successful processes performed much better. First, the Kafka producer currently doesn't wait for acknowledgments from the broker and sends messages as fast as the broker can handle. Data size is directly proportional to the performance time in producer performance.



	9.2 MB	13.9 MB	18.4 MB	36.8 MB
◆ Fail	170	230	280	2643
■ Success	6535	11470	4345	260

Figure 4. Consumer Performance on two servers

B. Experiment 2:Consumer Performance on two servers

Figure 4 shows a single consumer to consume various data sizes of messages 9.2 MB, 13.9 MB, 18.4 MB, 36.8 MB respectively. It configured the Kafka consumer to consume messages asynchronously by five batches size. Figure 4 shows the result. The x-axis represents the amount of data sent to the broker, and the y-axis corresponds to the total time in producer performance. It compares the total times on the difference between successful and failed processes with 40 partitions in various data sizes on two servers. The experiment emphasizes the total time by calculating based on consumer performance tools in Apache Kafka. The total time of failed processes is directly proportional to the data set. The comparison of successful process and failed processes are different .In 36.8 MB, the failed processing time raises than successful processes because of increasing messages lost. The system must reschedule processes for losing messages .Overall total time in this experiment is different in Experiment 1.

C. Experiment 3: Producer Performance on three servers

Figure 5 shows a single producer to publish various data sizes of messages 9.2 MB, 13.9 MB, 18.4 MB, 36.8 MB respectively. It configured the Kafka producer to send messages asynchronously by five batches size. Figure 5 shows the result. The x-axis represents the amount of data sent to the broker, and the y-axis corresponds to the total time in producer performance. It compares the total times on the difference between successful and failed processes with 40 partitions in various data sizes on three servers. The experiment emphasizes the total time by calculating based on producer performance tools in Apache Kafka. When we tested on three servers, the total time of failed processes raises than Experiment 1.The producer performance on three servers is more handle and faster than Experiment 1.

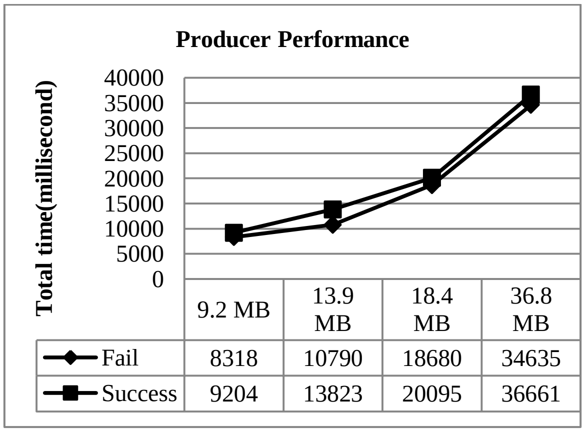


Figure 5. Producer Performance on three servers

D. Experiment 4: Consumer Performance on three servers

Figure 6 shows a single consumer to consume various data size of messages 9.2 MB, 13.9 MB, 18.4 MB, 36.8 MB respectively. It configured the Kafka consumer to receive messages asynchronously by five batches size. Figure 6 shows the result. The x-axis represents the amount of data sent to the broker, and the y-axis corresponds to the total time in consumer performance. It compares the total times on difference between successful and failed processes with 40 partitions in various data size on three servers. The experiment emphasizes the total time by calculating based on consumer performance tools in Apache Kafka. In Figure 6, the failed processing time of the most dataset is significantly rising than successful processes. More dataset is directly proportional to losing messages. The system spends more rescheduling time for failed processes. When we tested more data size on three servers, we need to handle consumer performance. When we tested more servers in Experiment 4, the processing time of the failed process is more increase than Experiment 2.

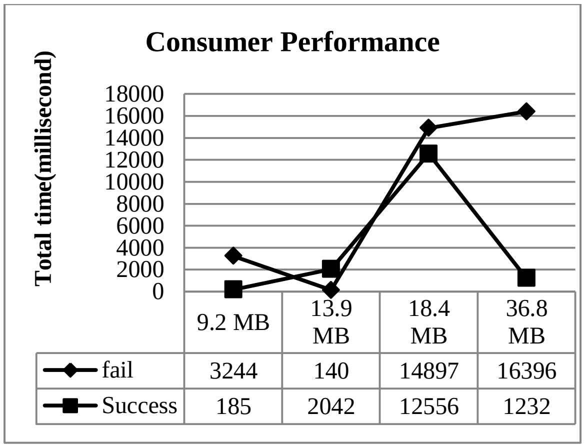


Figure 6. Consumer Performance on three servers

The system tested on two types of servers on the different number of datasets. We used performance tools in Apache Kafka to measure the performance of producer and consumer. The producer and consumer performance of experiments calculate based on Equation 1 to 4.The proposed system uses to input commands for measuring producer performance. There are broker lists, number of messages and topic name .Producer Performance commands show the result at the end of the performance testing. There is the start time of the test, the end time of the test, compression, message size, and batch size, total consumed messages bytes (MB), average consumed messages bytes per second, total consumed messages count, average consumed messages count per second. The system uses zookeeper host, number of messages and topic as input command for measuring consumer performance. The information included in the command output for consumer performance. There is the start time of the test, the end time of the test, total consumed messages bytes (MB), average consumed messages bytes per second, total consumed messages count, average consumed messages count per second.

In (1) to calculate Total consumed message bytes (MB) representα, β represents total bytes sent and memory size is 1024*1024.

$$\propto = \frac{(\beta * 1.0)}{(1024 * 1024)} \qquad (1)$$

In (2): Calculation the Average consumed MB bytes per second representα, βrepresents total messages sent and elapsed seconds.

$$\alpha = \frac{\beta}{\theta} \qquad (2)$$

In (3): Calculation the elapsed timeα,the end time (milliseconds) is β and the start time (milliseconds) is £.

$$\propto = \frac{(\beta - \text{£})}{1000.0} \qquad (3)$$

In (4): Calculation the Average consumed message bytes per secondα, β represents total messages sent and £ represents elapsed seconds.

$$\alpha = \frac{\beta}{\text{£}} \qquad (4)$$

Table 3. Experimental Summary for Producer

Data size (MB)	success /fail	No of messages per second on two Servers	No of messages per second on three Servers
9.2 MB	S	10940.2	12112.2
9.2 MB	F	33050.9	13402.4
13.9 MB	S	11227.8	12096.9
13.9 MB	F	51089.5	15497.3
18.4 MB	S	13194.2	11095.5
18.4 MB	F	41341.7	11936.1
36.8 MB	S	14596.6	12163.4
36.8 MB	F	44561.4	12874.9

Table 3 summarizes experimental results. The table compares the number of messages per one second between two servers and three servers. The results indicate that the number of messages per second on two types of the server is different. The number of messages tested on three servers is slightly different than two servers. The processing time of Figure (3) and (5) correlate to the number of messages per second in Table 3.The experimental results calculated based on Equation (1) to (4).According to the result, producer performance on three servers improves than two servers.

Table 4. Experimental Summary for Consumer

Data size (MB)	success /fail	No of messages per second on two Servers	No of messages per second on three Servers
9.2 MB	S	34117.9	1205194.6
9.2 MB	F	1287541	68171.4
13.9 MB	S	29157.1	163776.2
13.9 MB	F	1442182.6	2374692.9
18.4 MB	S	102626.2	35514.5
18.4 MB	F	1587103.5	29775.3
36.8 MB	S	1715096.2	1922068.9
36.8 MB	F	244064.5	54285.2

Table4 summarizes experimental results. The table shows the number of messages per one second on two and three servers respectively. The results indicate that the consumer performance can’t handle the messages and processing time. The numbers of messages are significantly different on the two types of servers. Some consumer consumes more messages in one second and some consumes fewer messages. The processing time of Figure (4) and (6) correlate to the number of messages per second in Table 4.The experimental results calculated based on Equation (1) to (4). According to the experiments, the system needs to control the performance on many servers. We need to reduce completion time and recover lost messages.

In the future, Message logging based check pointing will be used to solve the problem of lost messages and to reduce the completion time of the system. It is used to provide fault tolerance in distributed systems in which all inter-process communication is through messages. Message-logging protocols guarantee that upon recovery, no process is an orphan. This requirement can be enforced either by avoiding the creation of orphans during an execution. Check pointing is utilized to limit log size and recovery latency.

5. Conclusion

In this paper, the comparison of the processing time between the successful process and failed process on two types of the server was evaluated. The analysis of the results has shown that the comparison of four types of data set with servers. Any application that needs to process a large number of messages is tolerant lost a few. If the system handles lost messages, Kafka will get the more reliable messages. We conducted several experiments to determine higher performance and to get reliable streaming data. To develop the reliability of pipeline architecture, we need to evaluate the higher performance of processes in Apache Kafka.

As the future direction, we intend to handle the performance of Kafka processing in real time pipeline architecture by using Message logging based check pointing. By analyzing many processes, we can achieve more reliable data in Kafka-storm pipeline architecture.

6. References

[1] Jay Kreps,NehaNarkhede, Jun Rao,"Kafka: a Distributed Messaging System for Log Processing", May , 2015.

[2] Hassan Nazeer, Waheed Iqbal, Fawaz Bokhari,Faisal Bukhari, Shuja Ur Rehman Baig ,"Real-time Text Analytics Pipeline Using Open-source Big Data Tools" ,December ,2017.

[3] Martin Kleppmann, "Kafka, Semza and the Unix Philosophy of Distributed Data" Bulletin of the IEEE computer Society Technical Committee on Data Engineering, July, 2016.

[4] Paul Le Noac'h, Alexandru Costan , Luc Boug´e ," A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications" , 2017 IEEE International Conference on Big Data (BIGDATA), December ,2017.

[5] Wenjie Yang, Xingang Liu and Lan Zhang, "Big Data Real-time Processing Based on Storm", 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013.

[6] Mohit Maske, Dr. Prakash Prasad, "A Real Time Processing and Streaming of Wireless Network Data using Storm ", International Journal of Advanced Research in Computer Science and Software Engineering, January, 2015.

[7] SeverinSimko ,"Performance testing of Apache Storm framework", Masaryk University, Faculty of Informatics, Brno, Autumn 2015

[8] Muhammad Syafrudin 1 ID , Norma LatifFitriyani 1 ID , Donglai Li 1, GanjarAlfian 2 ID , Jongtae Rhee 1 and Yong-Shin Kang 3, "An Open Source-Based Real-Time Data Processing Architecture Framework for Manufacturing Sustainability" ,*Sustainability* Open Access Journal 2017 ,20 November 2017

[9] NishantGarg, "Apache Kafka", PACKT Publishing UK,October , 2013

[10] Ankit Jain, AnandNalya, "Learning Storm" PACKT Publishing UK, August, 2014.

[11] Tanenbaum Andrew, Tanenbaum-Distributed operating system, Wikipedia, p-100, 1994.

[12] http://zookeeper.apache.org/.

Image Processing

Face Detection using Fusion of Skin Detection and Viola-Jones Face Detection

Hla Myat Maw, K Zin Lin, Myat Thida Mon
University of Information Technology
Yangon, Myanmar
hmyatmaw@uit.edu.mm, kzinlin@uit.edu.mm, myattmon@uit.edu.mm

Abstract

Face Detection is the first step of any automated face and facial expression recognition system. It is applied in many applications such as computer vision and biometric applications like face recognition, expression recognition and classification, security access systems, video surveillance, and intelligent human computer interaction. A false detection rate and slow detection speed remain major problems in face detection. In this paper, we present a method to avoid these problems, by combining skin detection and Viola and Jones Face detection. Skin detection is used to reject false positive. After determining all possible face candidate regions, the Viola Jones detector is applied to detect face. Our method has high detection rate and accuracy than an individual face detector. It can perform quiet well in complex background and varying lighting conditions when it is applied images with complex backgrounds. Experimental results show the effectiveness of the proposed method.

Keywords-Face detection, Skin detection, Viola and Jones face detector, Color space model.

1. Introduction

Face detection is an essential step and usually in various computer vision and biometric applications such as face recognition, criminal investigation, and security access system, video surveillance, and human computer-interaction. Numerous researches were performed in the field of face detection and generally can be classified into four categories [1]: feature invariant approaches, template matching, knowledge based and appearance based methods. In the following, a brief review of these four categories is given. (1) Knowledge based methods are rule-based methods, which encode human knowledge about what a face is. For example, in our human mind, symmetric eyes, ears, nose and mouth are key face feature. Developing these methods in different situations is sometimes difficult because not all states are countable (2) Feature invariant approaches are grouping methods with aim to find robust structural features which are invariant to pose, lighting, etc. This method is one of the most important methods for face detection and features use low-level features such as intensity, color, edge, shape, and texture to locate facial features, and further, find out the face location. (3) Template matching methods compute the correlation between patterns of a face and an input image in order to detection. In these methods, the correlation of several patterns of face in different poses and the input images are stored to be a criterion for face validation. It is scale-dependent, rotation-dependent and computational expensive. (4)Appearance-based methods use models learned from training sets to represent the variability of facial appearance. Actually in these methods the templates are learned from face image samples. Generally, appearance-based methods utilize statistical analysis and machine learning to find characteristics related to face or non-face images.

Among feature-based face detection methods, skin color has been used and proven to be an effective feature to increase detection rate [3]. It is often considered to be a useful and discriminating image feature for facial area and usually employed as a preliminary step in face detection. It provides computationally effective yet robust to variation in scale, orientation and partial occlusion. Skin detection is also a challenging task since the skin color is sensitive to various factors including illumination, ethnicity, individual characteristics, subject appearances and camera characteristics. [4].

There are three primary steps for color-based skin detection in an image: representing the image pixels in a suitable color space, modeling the skin and non-skin pixels using an appropriate distribution and classifying the modeled distributions. A skin detection system utilizing skin color as a feature and HSV, YCbCr, YCgCr and YDbDr are the most appropriate color spaces for skin color detection.

The rest of this paper is organized as follows: Section 2 gives an overview of the related work. Section 3 presents background theory. The proposed system is presented in Section 4. Section5presents experimental result. Section 6 closes with a conclusion and future work.

2. Related work

A literature of some research work is presented to identify the best color model for a specific task. Fatma et al. [5] proposed technique is based on finding the maximum energy of histogram signal for skin which is limited to the ranges for each component of the color space under study. Different parameters such as energy of the histogram of each component of the color space, the limit of skin range in each color space and the maximum energy of the color spaces are used to evaluate. The result indicates that YCbCr provide better performance compared to RGB, YUV, HSV and CMYK color model. A detection rate of 97.51% was obtained using (Psychological Image Collection at Stirling) PICS database.

In [6] presents a human face detection scheme by combining a novel hybrid color models and Viola-Jones face detector. A hybrid skin color model RGB-CbCrCg was proposed for classifying skin and non-skin pixels. Afterward the segmented face regions are identified using Viola-Jones algorithm. A detection rate of 83.91% was obtained using (Edith Cowan University) ECU database.

In [7] describes a machine learning approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates. Viola and Jones face detector has become the standard to build successful face detection in real time, however, it produces a high false positive (detecting a face when there is none) and false negative rate (not detecting a face that's present) when directly applied to the input image. To deal with this problem, a various improvements have been proposed, such as using skin color filters (whether pre- filtering or post-filtering) to provide complementary information in color images. Though many experimental results have demonstrated the feasibility of combining skin color detection with Viola-Jones face detector to reduce false positive, both methods suffer from high false negative rate as some face regions may be ignored by detector. A detection rate of 76.1% was obtained using (Edith Cowan University) MIT + CMU database.

3. Background theory

In this section, various techniques which have been used in the proposed algorithm will be explained.

3.1. Haar-like features

Haar-like features are digital image features used in object recognition. They owe their name to their intuitive similarity with Haar wavelets and were used in the first real time face detector. Historically, working with only image intensities (i.e., the RGB pixel values at each pixel of image) lead to expensive computation in feature calculation. The idea of Haar-like feature is to consider the adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and compute the difference between these sums. This difference is then used as a feature response to categorize subsections of an image.

Figure 1 shows three kinds of rectangle Haar-like features. The value of a two-rectangle feature is the difference between the sums of the pixels within two rectangular regions. The regions have the same size and shapes are horizontally or vertically adjacent. A three rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally a four-rectangle feature computes the difference between diagonal pairs of rectangles.

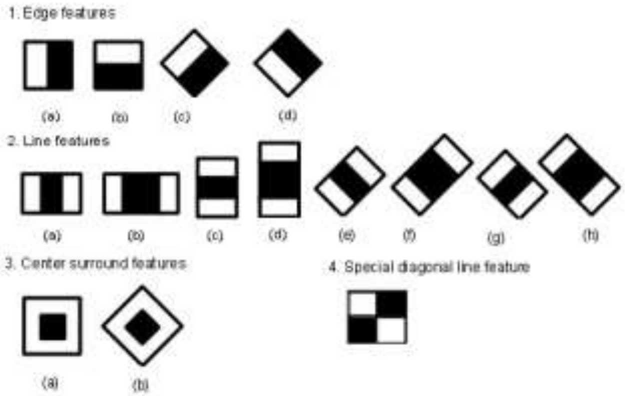


Figure1. Example rectangle features. The sums of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. Two-rectangle features are shown in (1), three-rectangle features are shown in (2), center surround feature are shown in (3) and a four rectangle features is shown in (4).

3.2. Integral image

Rectangle features can be computed very rapidly using the integral image [8]. The integral image at location x, y contains the sum of the pixels above and to the left of x, y Inclusive:

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x',y') \qquad (1)$$

Where $ii(x,y)$ is the integral image and $i(x,y)$ is theoriginal image.Using the following pair of recurrences:

$$s(x,y) = s(x,y-1) + i(x,y) \qquad (2)$$

$$ii(x,y) = ii(x-1,y) + s(x,y) \qquad (3)$$

Where s(x, y) is the cumulative row sum,
s(x,−1) = 0, and ii (−1, y) = 0, the integral image can be computed in one pass over the original image.

Using the integral image any rectangular sum can be computed in four array references (see Figure 2). We can easily see that the difference between two rectangular sums can be computed in eight references. Since the two-rectangle features defined above involve adjacent rectangular sums they can be computed in six array references, eight in the case of the three-rectangle features, and nine for four-rectangle features.

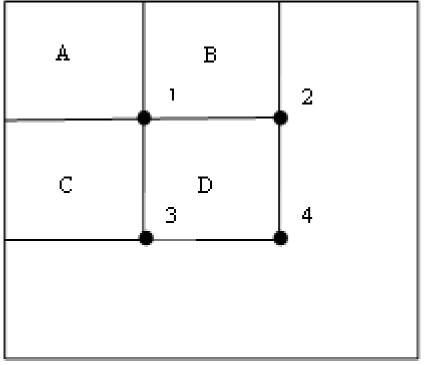


Figure2. The sum of the pixels within rectangle D can be computed with four array references. The value of the integral image at location 1 is the sum of the pixels in rectangle A. The value at location 2 is A+B, at location 3 is A+C, and at location 4 is A+B+C+D. The sum within D can be computed as 4+1-(2+3).

After calculation of a huge number of features for each analysis window, the AdaBoost algorithm is used for combining a small number of these features to form an effective classifier.

In continue the Adaboost classifier was used for feature classifying. This classifier has two functions. At the first, the appropriate feature has been selected for every template then these features will be learned. Generally, the purpose of the Adaboost algorithm is improving and increasing the efficiency of classified simple algorithms. The importance of the Adaboost algorithm is selecting a few numbers of these features according to the template structure. During the training process, these features are extracted for different training templates. Then those features with no change for all training data are selected. Some of the selected features, which create the most distinction between the two classes, are extracted as final features.

3.3. Skin color model

A color space is a geometrical representation of colors in a space and allows specification of colors through three components whose arithmetical values define an exact color. The native and most available representation of color images is the RGB color format describing the world view in three color matrices. Any other color space can be derived from a linear or nonlinear transformation from the RGB channels. Moreover, different skin color spaces may be combined to get improved. The existing color-space combination methods can be categorized as mixture of collections. Color space selection is the primary process in skin color modeling and further for classification. One or more color spaces can give an optimal threshold value for detection of pixels of skin in a given image. The choice or components are hybrid/fusion of color spaces.

Figure 3 shows YCbCr (Luminance, Chrominance) Color Model. YCbCr is an encoded non-linear RGB signal, commonly used by European television studios and for image compression work. As shown in fig. 3 color is represented by luma (which is luminance computed from non linear RGB) constructed as a weighted sum of RGB values [4]. YCbCr is a commonly used color space in digital video domain. Because the representation makes it easy to get rid of some redundant color information, it is used in image and video compression standards like JPEG, MPEG1, MPEG2 and MPEG4.Thetransformationsimplicity and explicit separation of luminance and chrominance components makes YCbCr color space [3]. In this format, luminance information is stored as a single component (Y), and chrominance information is stored as two color-difference components (Cb and Cr). Cb represents the difference between the blue component and a reference value. Cr represents the difference between the red component and a reference value. YCbCr values can be obtained from RGB color space according to eq. 4 to eq. 6.uses YCbCr space for skin detection.

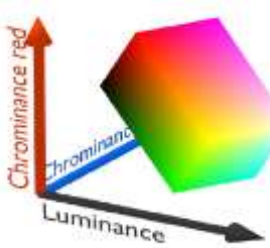


Figure 3. YCbCr Color model [12]

$$Y=0.299R+0.287G+0.11B \qquad (4)$$

$$Cr=R-Y \qquad (5)$$

$$Cb=B-Y \qquad (6)$$

3.4. Skin color detection

Skin detection is the process of finding skin-colored pixels and regions in an image or a video. This process is typically used as a preprocessing step to find regions that potentially have human faces and limbs in images. Skin detection techniques can be broadly classified as pixel-based techniques or region-based techniques. In the pixel-based skin detection, each pixel is classified as either skin or non-skin pixel individually depending on certain conditions. The skin detection based on color values is pixel-based. In region-based skin detection technique, spatial relationship of pixels is considered to define some area from given image as skin region.

The most important parameter which is used to detect skin pixels in the image is color. According to the importance of color in skin detection, the used color space is also important. HSV, YCbCr, HIS, HUV, and YIQ color spaces are persistent against light and luminance intensity changes. Eliminating the y component in these spaces reduces input space in addition to reduce dependency to light intensity. In color segmentation scheme needs a proper representation of color spaces to interpret image information in many cases. In this paper, we use YCrCb color space for detecting skin in color images.

4. Proposed system

The proposed system is fused skin detection based on YCbCr color model and face detection based on Viola and Jones algorithm for face detection from color images. Figure 4 shows the block diagram of the proposed system.

4.1. Skin detection

Skin color methodologies have been widely used in many applications. Skin color is an affective key for face detection since it provides computationally effective, robust information against rotation, scaling and partial occlusions. In general, the final goal of skin color detection is to build a decision rule, which will discriminate between skin and non-skin pixels. Skin color detection may fall into two main categories [9]: pixel-based skin detection methods and region-based skin detection methods. Pixel-based skin detection methods classify each pixel as skin or non-skin individually, independently from its neighbors. On the other hand, region- based skin detection methods try to take the spatial arrangement of skin pixels into account during the detection stage to enhance the methods performance.

In our proposed method, especially for real-time implementation, the pixel-based skin detection method is chosen for fast processing. We also use a color compensation step prior to skin detection, to reduce the effects of lighting. A pixel with color components (R, G, B) is detected as skin if the conditions given in (7) below hold. The second line in (7) ensures that RGB components must not be close together, which ensures grayness elimination. The third line in (7) ensures that R and G components must not be close together, which must be true for fair complexion.

(a) $R > 95 \& G > 40 \& B > 20$ and (7)
(b) $\max\{R,G,B\} - \min\{R,G,B\} > 15$ and
(c) $|R - G| > 15 \& R > G \& R > B$

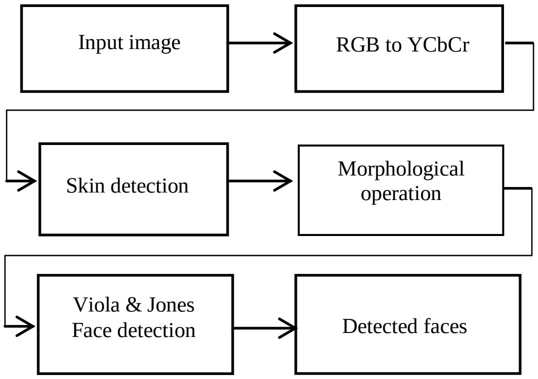
For YCbCr color system, a pixel (Y, Cb, Cr) is classified as skin if:
(a) $60<Y<255$ and (8)
(b) $100<Cb<125$ and
(c) $135<Cr<170$

4.2. Viola and Jones face detection

To classify the areas into faces or non faces, the Viola and Jones face detector is performed locally on all selected bounding boxes around connected pixel regions on an image. This allows help to decrease the false positives in face detection. Viola & Jones [7] have presented a face detection method based on an over-complete set of Haar-like features in (Figure.1) which are calculated in scaled analysis windows. The rectangular Haar-like features are sensitive to edges, bars and other similar structures in the image and they are computed using an efficient method based on the integral image concept.

After calculation of a huge number of features for each analysis window, the AdaBoost algorithm is used for combining a small number of these features to form an effective classifier.

Thus, if the image passes through all the stages then it is a face. If it fails even in any one of the stages then it is not a face. This gives an overall description of the algorithm.


Input image

RGB to YCbCr

Skin detection

Morphological operation

Viola & Jones Face detection

Detected faces

Figure 4. Block diagram of the proposed system

The steps of implemented algorithm are:

Step1: Read the image, and capture the dimensions.
Figure 5 shows the original image.

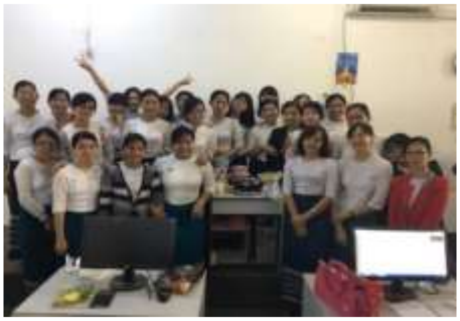


Figure5. Original Image

Step2: Lighting compensation

Lighting compensation is used for reducing the effects of lighting. Figure 6 shows the results of lighting compensation.

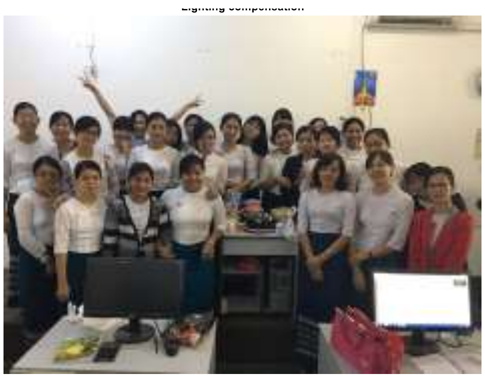


Figure6. Lighting Compensated Result

Step3: Skin region extraction

Applying the color segmentation to the image shown in Figure 6 using inequalities and the result is shown in Figure 7. The output image is a binary image that contains ones in the skin regions and zeros in non-skin regions. The detected skin regions may be discontinuous; this discontinuity may be due to lighting effects that leads to missed skin pixels or due to the presence of non-skin face features like the eyes and brows.

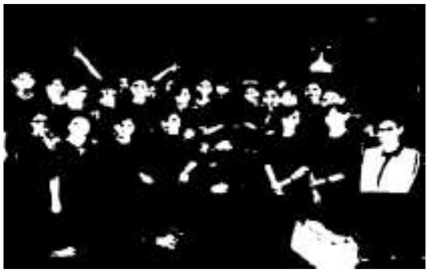


Figure7. Extracted Skin Result

Step4: Noise removing

In this step, morphological operation is applied to reduce false positives. Figure 8 shows the output after noise removing.

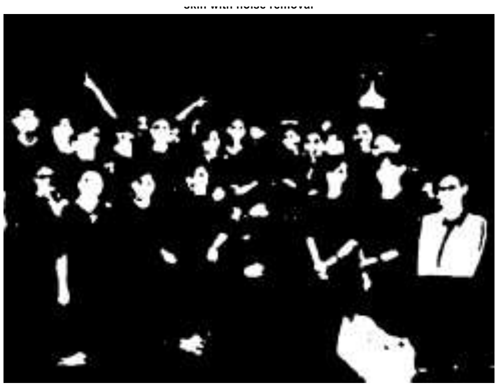


Figure8. De-noised Skin Result

Step5: Find skin color block

Figure 9 shows the result of skin color block.

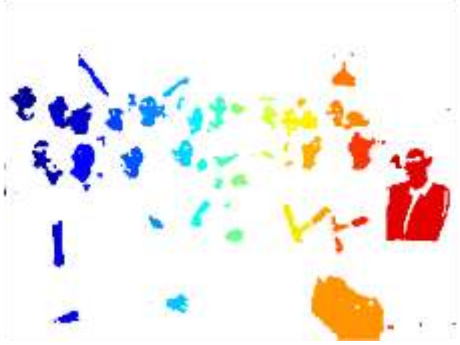


Figure9. Skin Color Block Result

Step6:.Face Detection

The Viola and Jones face detector is used to detect face from skin region. The algorithm is to scan a sub window capable of detecting faces across the skin region using features consisting of two or more rectangles. The result is shown in figure 10.

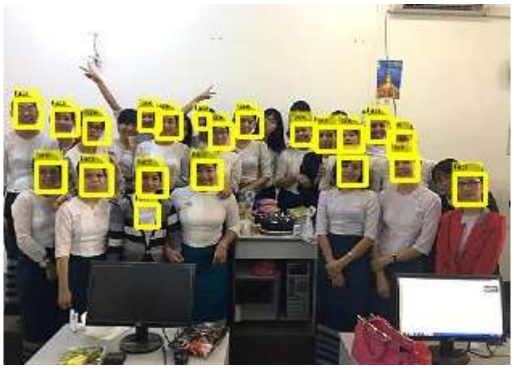


Figure10. Detected Face(s) Result

5. Experimental results

In order to evaluate the performance of the proposed method, many experiments have been carried out using a total of 30 images of varying lighting conditions and complex background.

We also compare our proposed method with Viola-Jones face detector as an individual method. The comparison is based on the accuracy, the false positives rate, and the false negatives rate, which are defined below Table1.

The results presented in Table 1 were obtained using a variety of face dataset. The analysis of these results show clearly, that the accuracy of the Viola & Jones detector increases when a skin color detection is used as a prior stages which means that this face detector gives low false negatives rate when it is not applied directly to the entire input image. The proposed system improves significantly the accuracy by overcoming the problem of false detection. The faces of those people wearing caps, glasses are also detected.

Table. 1. Comparisons of our results with other face detector method

	Accuracy	FNR	FPR
Viola & Jones [07]	79.46%	20.54%	34%
Proposed system	86.55%	0.89%	0.15%

- False Positives Rate (FPR) = the ratio of the number of detected false positives to the total number of faces.
- False Negatives Rate (FNR) = the ratio of the number of false negatives to the total number of faces.
- Accuracy = 1-(FPR+FNR)/2

6. Conclusion and future work

In this paper, we proposed system using a fusion of skin detection and Viola and Jones face detector to improve the accuracy of face detection in varying illumination and complex background. Skin detection is used before face detection to reduce false negative rate and overcome variation in pose and illumination. The experimental results presented illustrate the effectiveness of this method, compared to some other method proposed in the literature, especially the well- known Viola &Jones face detector. The proposed method works for non-frontal face and frontal face and different lighting condition. But it takes times for complex background.

In future, more research work should continually focus on human face detection for people of different races and computation time should be further saved for real time applications.

7. References

[1] M.-H. Yang, D. J. Kriegman, and N. Ahuja, ‟Detecting Faces in Images: A Survey”, IEEE Transaction on Pattern Analysis and Machine Learning, vol. 24, no.1, 2002, pp. 34-58.

[2] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, “Face Detection in Color Images”, IEEE Transaction on Pattern Analysis and Machine Learning, vol. 24, no.5, 2002, pp. 696-706.

[3] Z. S. Tabatabaie, R. W. Rahmat, N. I. B. Udzir and E.Kheirkhah, “A Hybrid Face Detection System using Combination of Appearance-based and Feature-based Methods”, International Journal of Computer Science and Network Security, vol. 9, no.5, 2009, pp. 181-185.

[4] Y. Wang, B. Yuan, “A Novel Approach for Human Face Detection from Color Images under Complex Background”, Pattern Recognition, vol. 34, no.10, 2001, pp. 983-1992.

[5] FatmaSusilawatiMohamad, Abdulganiyu Abdu Yusuf, ZahraddeenSufyanu, “Evaluation of Suitable Color Model for Human Face Detection”, International Conference on Electrical Engineering and Electronics Communication System (ICEEECS), 2015.

[6] Guan-Chunluh, “Face Detection Using Combination of Skin Color Pixel Detection and Viola-Jones Face Detector”, International Conference on Machine Learning and Cybernetics, Lanzhou, 13-16 July, IEEE , 2014.

[7] Viola Paul, and Michael Jones. "Rapid Object Detection using A Boosted Cascade of Simple Features." Proceedings of the 2001 IEEE Computer Society Conference on. Computer Vision and Pattern Recognition, CVPR 2001.

[8] P. Viola and M. J. Jones, “Robust Real Time Face Detection”, International Journal of Computer Vision, vol. 57, no. 2, 2004,pp. 137-154.

[9] Deepak Ghimire and Joonwhoan Lee, “A Robust Face Detection Method Based on Skin Color and Edges”, J. Inf. Process Syst., vol.9, no.1, March 2013.

[10] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, “A Survey of Skin-Color Modeling and Detection Methods, Pattern Recognition”, vol.40, no.3, pp.1106-1122, 2007.

[11] Amit Pal, “Multicues Face Detection in Complex Background for Frontal Faces”, International Machine Vision and Image Processing Conference, IEEE 2008.

[12] https://en.wikipedia.org/wiki/Talk:YCbCr

Object-based Urban Land Use Classification using Deep Belief Network

Su Wai Tun, Khin Mo Mo Tun
University of Information Technology
Yangon, Myanmar
suwaitun@uit.edu.mm, suwaidun@gmail.com
khinmomotun@uit.edu.mm, khinmo2htun@gmail.com

Abstract

Urban land use information is very important for urban planning, regional administration and management. Classification of urban land use from high resolution images remains a challenging task, due to the extreme difficulties in differentiating complex spatial patterns to derive high-level semantic labels. Deep learning is a powerful state-of-the-art technique for image processing including remote sensing images. The Deep Belief Networks (DBN) model is a widely investigated and deployed deep learning architecture. It combines the advantages of unsupervised and supervised learning and can archive good classification performance. In this paper, deep belief network model is used to improve the performance of object-based land use classification. First, to achieve an object-based image representation, the original image is segmented into objects by graph-based minimal-spanning-tree segmentation algorithm. Second, spectral, spatial and texture features for each object are extracted. Then all features are put into deep belief network and the parameters of the network using training samples are trained. Finally, all objects are classified by network.

Keywords- Classification, Deep Belief Network, Land Use

1. Introduction

Land use information is essential for urban planning and management and also provides a key input to urban and transportation models, and is essential to understanding the complex interactions between human activities and environmental change. In recent years, high resolution remote sensed images from satellites, planes, and unmanned aerial vehicle (UAV) have been widely used for classification. Information on urban land use within high resolution images is presented implicitly as patterns or high level semantic functions, in which some identical low-level ground features or object classes are frequently shared amongst different land use categories. This complexity and diversity of spatial and structural patterns in urban areas makes its classification into land use classes a challenging task. It is very important to develop robust and accurate urban land use classification techniques by effectively representing the spatial patterns or structures lying in high resolution remotely sensed data. Pixel based image analysis (PBIA) has been a popular method to classify remote sensed images given its simplicity and high efficiency. The PBIA method cannot take full advantage of the texture/contextual information found in high resolution images thus the results display a salt and pepper effect after classification. Because of this problem, object-based image analysis (OBIA) has become a main method in land-use/land-cover (LULC) applications over the last decade. OBIA was presented to overcome the drawbacks of PBIA when classifying high resolution image.

Recently, deep learning has become the new hot topic in machine learning and pattern recognition, where the most representative and discriminative features are learnt end-to-end, hierarchically [6]. The DBN employs a hierarchical structure with multiple stacked restricted Boltzmann machines and works through a layer by layer successive learning process. This paper presents object-based land use classification based on deep belief network (DBN) to improve classification result.

2. Related Works

In literature, [1] land use classification method based on stack autoencoder has been proposed by Anzi Ding, Xinmin Zhou. This method is tested in GF-1 images with 4 spectral bands and spatial resolution of 8 m. They show that the method based on SAE is more accurate in classification result than support vector machine and back propagation neural network.

[2] Dino Ienco, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel have proposed land cover classification method based on Deep recurrent neural networks. This proposed model has validated on two different data set showing that this framework efficiently deals with both pixel- and object-based classifications.[4]Deep convolutional neural network for land-cover classification method has proposed by Grant J. Scott, R. England, William A. Starms, Richard A. Marcum and Curt H. Davis.

[5] Chen and Guestrin published a new, regularized implementation of gradient boosting machines (GBM), called extreme gradient boosting classifier (Xgboost). It has made a very strong impact on the machine learning community, being the winning solution of most machine learning competitions. Then, Stefanos Georganos, Tais Grippa, Sabine Vahuysse, Moritz Lennert, Michal Shimoni evaluated the implementation of Xgboost for very high resolution object-based land use land cover urban classification. The results demonstrated that optimized Xgboost with a Bayesian model consistently outperforms random forest (RF) and support vector machines (SVMs) in different very high resolution data sets and classification schemes but at the cost of increased computational time.

[6]Ce Zhang, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare proposed object-based convolutional neural network (OCNN) for urban land use classification. Their proposed method starts with an initial image segmentation to achieve an object-based image representation. They used Mean-shift segmentation ,as a nonparametric clustering approach, to partition the image into objects with homogeneous spectral and spatial information. Then they developed two CNN networks with different model structures and window sizes to predict linearly shaped objects (e.g. Highway, Canal) and general (other non-linearly shaped) objects. Then a rule-based decision fusion was performed to integrate the class-specific classification results. Their proposed OCNN method was tested on aerial photography of two large urban scenes in Southampton and Manchester in Great Britain. The classification accuracy and computational efficiency of their method outperformed the Pixel-wise CNN, contextual-based MRF and object-based object-based image analysis SVM methods.

[10]Qi Lv, Yong Dou, Xin Niu, Jiaqing Xu, Jinbo Xu,and Fei Xia proposed a classification approach based on the DBN model for detailed urban mapping using polarimetric synthetic aperture radar (PolSAR) data. Through the DBN model, effective contextual mapping features can be automatically extracted from the PolSAR data to improve the classification performance. Two-date high-resolution RADARSAT-2 PolSAR data over the Great Toronto Area were used for evaluation. Their DBN-based method outperformed support vector machine (SVM), conventional neural networks (NN), and stochastic Expectation-Maximization (SEM) and produces homogenous mapping results with preserved shape details.

3. Theory Background

In machine learning, a deep belief network (DBN) is a generative graphical model, or alternatively a class of deep neural network, composed of multiple layers of latent variables ("hidden units"), with connections between the layers but not between units within each layer. When trained on a set of examples without supervision, a DBN can learn to probabilistically reconstruct its inputs. The layers then act as feature detectors. After this learning step, a DBN can be further trained with supervision to perform classification.

Restricted Boltzmann Machine, unsupervised learning, has the advantage of fitting the feature of the samples. So when we have an output of the hidden layer in a RBM, we can use it as the visible layer’s input of another RBM. This process can be regard as further feature extraction from the extracted feature of our samples. With this kind of thought, Hinton raised Deep Belief Network (DBN) in 2006, which is based on RBM. As the Figure 1 shows, by using the output of the upper RBM’s hidden layer as the input of the lower RBM’s visible layer, we get a Deep Belief Network. This DBN is stacked by three RBMs.

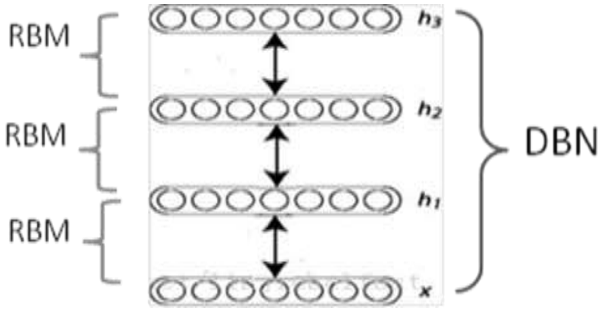


Figure 1. A DBN stacked by three RBMs

4. Object-based classification based on Deep Belief Network

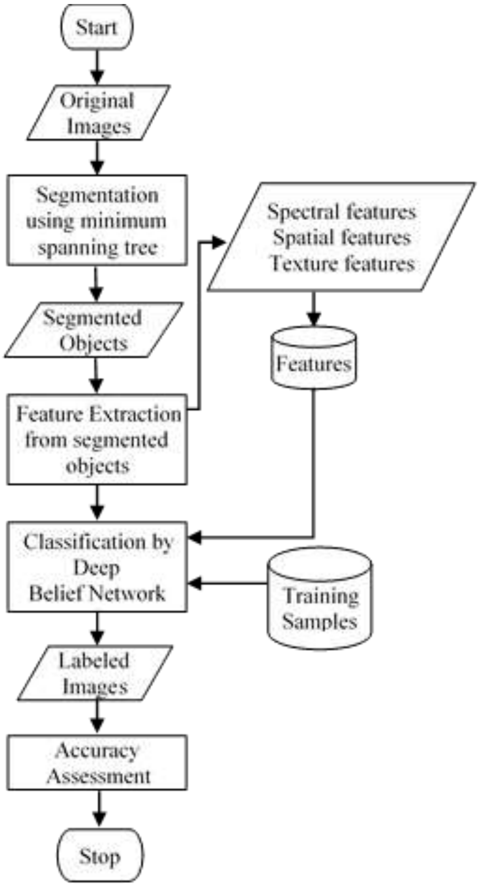


Figure 2. System Overview

4.1. Graph-based Minimal Spanning Tree Segmentation

Image segmentation is a process of partitioning a raster image into multiple segments. Many segmentation algorithms have been proposed, such as watershed, level set , etc. In this paper, a graph-based minimal-spanning-tree segmentation (GBS) method is used. The main concept in GBS is that each pixel of an image is considered as a vertex of an undirected graph, and a four neighborhood adjacent pixel-pair as an edge. The weight of each edge is calculated by the dissimilarity between two vertexes using the dissimilarity function. Then, all edges must be inserted into a minimal-spanning tree (or minimum-spanning-forest)in ascending order of its weight. The whole insertion progress is the merge-process and each tree represents an object. When the weight of the trees to be merged is smaller than the threshold, the merge-process will stop. At the end, some post segmentation procedures are implemented. Once the minimal-spanning-tree(or minimum-spanningforest) is formed, there always exist small regions (objects) in it. These are called "silver objects," and could be eliminated by forcibly merging them into their largest neighboring object. Another significant operation after segmentation is to convert results from a labeled raster image into vector data format (such as ESRI Shapefile format), referred to "Polygonization" (or "Vectorization"). The scale of the segmentation result is decided by a threshold, but it is hard to select it robustly .Therefore a trial and error method is used to get an appropriate segmentation scale.

4.2. Feature Extraction

To identify the category of an object automatically by supervised classification method, the features of object should be extracted. Features of an object are calculated by all the pixels and shape (contour) of an object. Three types of features are taken into consideration: spectral feature, spatial features, and texture features. They will be separately described in three tables.

1) Spectral features: Spectral features are the statistical attributes of an object found in the spectral bands of an image. A spectral feature can be an attribute of a single band (such as mean-value of the band) or all bands (such as brightness). All spectral features taken into consideration are listed in Table 1.

2) Spatial features: Spatial features depict the position and geometry information for an object, and calculated from the contour of the object polygonized from the pixels. All spatial features are listed in Table 2.

3) Texture features: Texture features include texture information of all pixels within the object. Image-texture refers to particular frequencies of change in tones and their resulting spatial arrangements. Haralick features for gray level co-occurrence matrix (GLCM) and gray level difference vector (GLDV) is used as the texture feature as listed in Table 3.

Table 1. Spectral Features

Name	Target	Description
Brightness	All bands	Mean value of all an object's pixels in all of a digital numbers spectral bands.
Mean	Single band	Mean value of an object's all pixels' digital numbers
Standard Deviation	Single band	Standard deviation of an object's all pixels' digital numbers
Max	Single band	Max value of an object's all pixels' digital numbers
Min	Single band	Min value of an object's all pixels' digital numbers
Mean of inner border	Single band	Mean value of pixels' digital numbers on the inner border of an object
Skewness	Single band	The Skewness feature describes the distribution of all the image layer intensity values of all pixels that form an object. A normal distribution has a skewness of zero

Table 2. Spatial Features

Name	Target	Description
Area	Contour	The area of the object's region
Asymmetry	Contour	The Asymmetry feature describes the relative length of an image object, compared to a regular polygon
Border Index	Contour	The border index feature describes how jagged an image object is; the more jagged, the higher its border index
Border Length	Contour	Border Length of an image object's region
Compactness	Contour	The Compactness feature describes how compact an image object is
Elliptic Fit	Contour	The elliptic fit feature describes how well an image object fits into an ellipse
Elongation	Contour	Elongation describes the ratio of the long side and the short side of the minimum-bounding-rectangle of an object
Radius of largest enclosed ellipse	Contour	The radius of largest enclosed ellipse of the object
enclosed ellipse		
Radius of smallest enclosing ellipse	Contour	The radius of smallest enclosing ellipse of the object
Rectangular Fit	Contour	The similarity of an image object to a rectangle
Roundness	Contour	The similarity of an image object to a circle
Shape Index	Contour	The smoothness of an image object border
X Center	Contour	X-coordinate of the center of an object
X Max	Contour	Maximum x-coordinate of an object
X Min	Contour	Minimum x-coordinate of an object
Y Center	Contour	Y-coordinate of the center of an object
Y Max	Contour	Maximum y-coordinate of an object
Y Min	Contour	Minimum y-coordinate of an object

Table 3. Texture Features

Name	Target	Description
Homogeneity	GLCM	The GLCM is a tabulation of how often different combinations of pixel gray levels occur in a scene. A different co-occurrence matrix exist for each spatial relationship
Contrast		
Dissimilarity		
Entropy		
Ang. 2nd Moment		
Mean		
Standard Deviation		
Correlation		
Ang. 2nd Moment	GLDV	The GLDV is the sum of the diagonals of the GLCM. It counts the occurrence of references to the neighbor pixels' absolute differences.
Entropy		
Mean		
Contrast		

4.3 Classification of Deep Belief Network

The DBN architecture and the general methodology included in DBN machine learning are described in this section. DBN is a multilayered architecture that consists of one visible layer and multiple hidden layers. The visible layer of a DBN accepts the input data and transfers the data to the hidden layers to complete the learning process [9].

Algorithm 1 Deep Belief Network Model

```
Input : Input Data $D$, Maximum number of layers $ML$,
        Number of Neuron for Each Layer $N$,
        Maximum number of Epochs $ME$
   Initialize $D, ML, N, ME$
        For Layer = 1: $ME$   do
            Train Network using RBN learning rule
            Save weights of hidden-visible connections
            and biases
        End For
   Back Propagation Classification
Output: Labeled Data
```

The overall learning process of DBN model is described in algorithm 1. As shown in algorithm 1, the input data, total number of hidden neurons in each hidden layer and maximum number of epochs for the model training process are required and initialized before the start of the DBN training process. Each layer of DBN is trained using the restricted Boltzmann machines (RBM) learning rule with two learning steps: positive and negative phases. In positive learning phase, data are transferred from bottom visible layer to hidden layer and the probability of generating hidden units are determined as $p(h\backslash v, W)$. In negative phase, a reconstruction of the data from previous visible layer are operated and the probability of generating visible units are determined as $p(v\backslash h, W)$. The data vector is used as an input to the visible units. After a maximum number of training epochs, the repetitive positive and negative phases in the training of RBM layers will result in trained weights and generated visible units. The overall system function of DBN learning procedure can be expressed as the probability of generating a visible vector (v) as a function of weights and hidden vectors based on RBM learning rule [8]. The probability of generating a visible vector p(v) by DBN learning process can be formulated using the probability of generating visible units in the reconstruction phase of the previous epoch and the probability of generating hidden units in the positive phase of the current epoch as

$$p(v) = \sum_{h} p(h\backslash v, W) p(v\backslash h, W) \qquad (1)$$

An iterative process from a lower layer to a higher layer continues till the maximum number of layers is trained. Each RBM is individually trained and the weights and biases are saved during the DBN training process. At the end of the training process, the data is transferred from bottom visible layer (data layer) to higher invisible layers throughout the DBN architecture [9].The DBN layer by layer training is an unsupervised learning process that cannot provide class labels of the training data. The label information of the training data will be used during the back-propagation training.

4.3.1. Stacked RBM learning. DBN is constructed with stacked RBMs. Training of the DBN model is completed through training of each RBN structure using RBN learning rule. Each RBN unit consists of two layers. There are

a number of neurons in each layer and there is no synaptic weight connection between neurons within the same layer [7].

Algorithm 2 Stacked RBM Model

```
Input: Input Data D, Maximum Number of Epochs ME,
       Number of Hidden Layers and Batches Numb
  Initialize Symmetric Weight and Biases
    For Batch =1: Numb do
      For Epoch =1: ME do
        Learn Positive Phase using Eq : (2)
        $P(h_j = 1\backslash v) = sigm(-b_j - \sum_k v_k w_{jk})$
        Learn Negative Phase using Eq: (3)
        $P(v_k = 1\backslash h) = sigm(-b_k - \sum_j h_j w_{jk})$
        If Epoch < 5, then
           Momemtum = Initial momentum
        Else
           Momemtum = Final momentum
        End If
        Update Weights and Biases
      End For
    End For
```

The iterative learning process for one RBN unit is described in algorithm 2. In this algorithm, the synaptic weights and biases of all neurons in each RBN layer are initialized at the beginning of the training process. After that, the RBN unit will be trained repeatedly with input training data. The training dataset is often divided into mini-batches with a small number of data vectors and weights are updated after treating each mini-batch. Each training epoch consists of two phases, the positive phase and negative phase. The positive phase transforms the input data from visible layer to the hidden layer. In negative phase, a reconstruction of the neurons of the previous visible layer is operated. The positive phase of RBM learning can be denoted mathematically as

$$P(h_j = 1\backslash v) = sigm(-b_j - \sum_{k} v_k w_{jk}) \qquad (2)$$

During the states of neurons in visible layer neurons in visible layer are reconstructed, the negative phase can be denoted mathematically as

$$P(v_k = 1\backslash h) = sigm(-b_k - \sum_{j} h_j w_{jk}) \qquad (3)$$

where h_j and v_k are the states for the jth neuron in hidden layer and the kth neuron in the visible layer respectively. The visible and hidden layer neurons are binary stochastic neurons with binary states 0 or 1, which representing on and off conditions of the neurons in the learning process.

After learning process for both positive and negative phases, synaptic weights and biases can be updated based on state vectors of neurons in both hidden layer and visible layers [8]. The update of synaptic weight, w_{jk}, can be denoted as

$$\Delta w_{jk} = \delta((v_k h_j)_{data} (v_k h_j)_{recon}) \qquad (4)$$

where δ is a value between 0 and 1, denoting the learning rate; $(v_k h_j)_{data}$ is the pairwise product of the state vectors for the jth neuron in the hidden layer and the kth neuron in the visible layer after positive phase learning process whereas $(v_k h_j)_{recon}$ denotes the pairwise product after the negative phase learning process for reconstruction of the visible layer. The same learning rule is utilize for bias updating, but individual hidden and visible units are used instead of pairwise products [8].To stabilize the RBN learning process, a momentum is usually used in updating the synaptic weights and biases. With momentum, the weight update,Δw_{jk} , at the current epoch and formulated as

$$[\Delta w_{jk}]_n = \left(m[\Delta w_{jk}]_{n-1}\right) + \delta((v_k h_j)_{data} - (v_k h_j)_{recon}) \quad (5)$$

The initial and final momentums utilized in the RBM training process are 0.5 and 0.9 respectively [8]. The learning parameters such as weights and biases of each RBM in the DBN model will be continuously optimized until a maximum number of training epochs are reached. This completes the training of one RBM and the process will be continued until all RBMs in the DBN structure are trained.

4.3.2. Back-propagation learning. After layer by layer learning process, the next step of the DBN training is the supervised learning process that will be completed by the back-propagation training algorithm. The supervised learning use labeled data for the training of the DBN model. Unlike the unsupervised DBN training process considers all DBN layers simultaneously. The back-propagation training is continued until the network output reaches the maximum number of epochs. After the supervised back-propagation training process, the trained DBN model can be further fine-tuned to improve classification accuracy through fine-tuning algorithms.

5. Datasets and Experimental Result

Experiment is conducted on World View-2 dataset including multispectral image which was obtained on January 13, 2010. The spatial resolution of multispectral image is 1.8 meter. The area is located in the Xihu District Hangzhou, China at30°14052.34@N, 120°6016.77@ E, covering an area of approximately 312.43 km. The dataset contains three spectral bands, which represents the red, green, blue band separately. The task in the experiments was to classify all pixels in image into six categories: water, bare land, vegetable, buildings, road, shadow. Part of the training datasets are shown in Fig 3 .

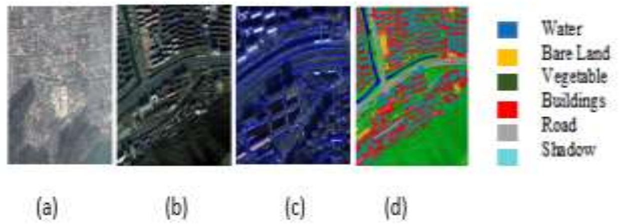


Figure 3. (a) The original image of World View-2 (b) Preprocessed image (c) Objects generated by GBS algorithm (c) Training samples with the original image (d)Training samples without the original image (part of the sample)

In the experiments, preprocessing and segmentation of the image are implemented sequentially as shown in Fig 3. Then 181 features (19 spectral features, 18 spatial features and 144 texture features) are generated. Finally, objects are classified by the deep belief network, which was trained using training dataset and testing dataset. In this experiment, the proposed object-based classification approach based on DBN is compared with other approaches: Bayes (Naïve Bayes) and linear support vector machine (linearSVM). Most procedures are the same to object-based classification approach using DBN except classification part. The overall accuracy of different methods are shown in Fig 4.

Table 4. overall accuracy of Naïve Bayes, Linear SVM, DBN

Overall accuracy (%)	
Naïve Bayes	68
Linear SVM	71
DBN	77.5

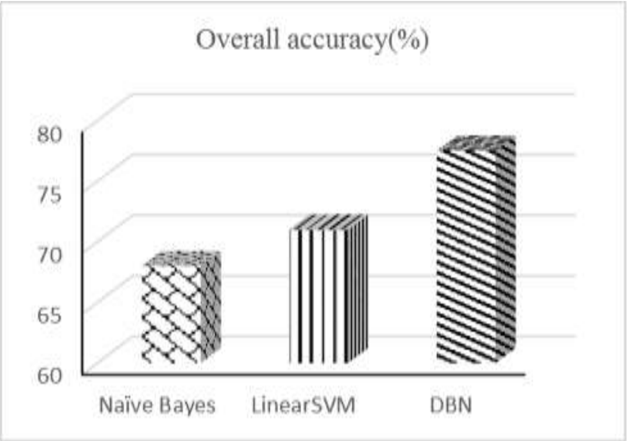


Figure 4. Object-based classification method based on DBN compared to other classification methods

To evaluate the effectiveness and generalization of the proposed classification approach based on DBN, the same experiment on the World View-2 dataset. The overall accuracy of each classifier as shown in table and figure which suggest that proposed classification approach based on DBN is more accurate than Bayes (Naïve Bayes) and linear support vector machine(linearSVM) on the World View-2 dataset.

6. Conclusion

This paper presented a object-based land use classification approach based on DBN. The results demonstrate that classification approach based on DBN outperforms Bayes and LinearSVM. In the existing object-based land use classification approach, the DBN model has not been used yet. Therefore, to improve the accuracy of the object-based land use classification approach, the DBN is used.

7. References

[1]Anzi Ding, Xinmin Zhou, “Land-use Classification with Remote Sensing Image Based on Stacked Autoencoder”, International Conference on Industrial Informatics, 2016.

[2]Dino Ienco, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel, “Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks”, IEEE Geoscience and Remote Sensing Letters, 2017.

[3]Grant J.Scott, Matthew R. England, William A. Strams, Richard A.Marcum and Curt H. Davis, “ Training Deep Convolutional Neural Networks for Land-Cover Classification of High-Resolution Imagery”, IEEE.

[4]Liangpei Zhang, Lefei Zhang, Bo Du, “ Deep Learning for Remote Sensing Data” , IEEE Geoscience and Remote Sensing Magazine” , 2016.

[5] Stefanos Georganos, Tais Grippa, Sabine Vahuysse, Moritz Lennert, Michal Shimoni , “ Very High Resolution
Object-Based Land Use-Land Cover Urban Classification Using Extreme Gradient Boosting”, IEEE Geoscience and Remote Sensing Letters, 2018.

[6] Ce Zhang, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, “ An Object-based Convolutional Neural Network (OCNN) for Urban Land Use Classification” , 2018.

[7] Lee H, Grosse R, Ranganath R, Ng AY, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations “ International conference on machine learning. Montreal, Canada, 2009.

[8] Hinton GE. “A practical guide to training restricted Boltzmann machines. Momentum”, 2010.

[9]Hinton GE, Osindero S, Teh YW.” A Fast learning algorithm for deep belief nets”, Neural Computation, 2006.

[10]Qi Lv, Yong Dou, XIn Niu, Jiaqing Xu,Jinbo Xu and Fei Xia, “ Urban Land Use and Land Cover Classification Using Remotely Sensed SAR Data through Deep Belief Networks ”, Journal of Sensor, 2015.

Adaptive Morphological Operation for High-Performance Weather Image Processing

Swe Swe Aung, Itaru Nagayama, Shiro Tamaki
Department of Information Engineering, University of the Ryukyus
sweswe@ie.u-ryukyu.ac.jp,nagayama@ie.u-ryukyu.ac.jp,shiro@ie.u-ryukyu.ac.jp

Abstract

Morphological operations have been an integral part of enhancement of digital imaging programs, especially for filtering noise for improving the quality of image by utilizing the two most basic morphological operations, named as erosion and dilation, altogether. The main role of dilation is to fill the defined region in an image with pixels, while erosion removes pixels from the region. As we know, the method of erosion followed by dilation or dilation followed by erosion is indeed an attractive approach amongst researchers to deal with filtering noise problems. However, this approach needs more computation time and has a high percentage of losing essential pixel area. To cover these issues, this paper introduces a new approach called adaptive morphological operation to boost the performance of image enhancement. Based on 2011, 2013, 2015, and 2016 weather image datasets collected from WITH radar, which is installed on the rooftop of Information Engineering building, University of the Ryukyus, the experimental results confirm that the proposed approach is more efficient than the conventional approach.

Keywords-Adaptive Morphological Operation, Dilation, Erosion

1. Introduction

It is never possible to be able to collect a perfect real dataset due to data corruption because of sensor or acquisition devices, and data transmission. Simply, the corruption acts like diseases that gradually swallows human's life. Likewise, the data corruption constantly forces algorithms struggle with prediction or classification work and face performance degradation in terms of mainly prediction accuracy. Thus, any kind of noise is inescapable in data collection and data preparation processing for the next advanced processes.

Thus, in the case of rainfall radar images, the shape of the rainfall region in the radar images tends to be variable and is easily covered with noise. Therefore, it is difficult to apply a fixed method to detect the rainfall region. It is necessary to develop a new method that can be applied to meteorological images that vary spatiotemporally.

Morphological filters have been using as a powerful tool for removing noises, shape detection, boundary detection, etc., by applying the two most basic approaches (erosion and dilation). In more details, erosion removes pixels from the predefined region in an image, while dilation fills that region with pixels. These two operations occupy the completely inverse relations. In the way, erosion followed by dilation or dilation followed by erosion is kind of dual operation in noise filtering problems. According to our experimentations, this dual process still lacks maintaining the indispensable pixels with absorbing high computational time.

For these issues, in the work of this paper, we focus on not only the reduction in noise but also in computational time of morphological approach by assigning the appropriate morphological operation (dilation or erosion) based on the adjustment of the local pixel density in an image, instead of applying two operations directly to that region. This simple and intuitive concept is named as an adaptive morphological operation.

The process by these two different operations, a sense comes to us is that if the pixel density of the region is high, then dilation operation is selected for that region. Otherwise, erosion has to take the responsibility for that region. Therefore, the new adaptation approach measures the pixel density of a targeted region before applying one of those two operations to that region. By doing so, this new approach reduces computational time obviously and fairly prevents the important pixels from wearing away.

2. Related Works

Morphological filters are the central theme of image enhancement, such as noise reduction, shape detection, etc. The authors of [1] proposed a system that utilized morphological operation for removing the salt and pepper noise from the input image with different structuring elements. The main methods of morphological filters that are erosion followed by dilation and dilation followed by erosion are applied to remove this kind of impulsive noise.

The authors of [2] primarily studied mathematical morphological operations such as dilation, erosion, opening, closing, fill and majority operations, which were used to accomplish filtering noise and enhance the appearance of binary images. As reported by the experimental results, they proved that noise could be effectively removed from binary images using combinations of erode-dilate operations.

Likewise, reference [3] tackled the problems of speckle noise removal and edge detection using the basic operations of mathematical morphology (dilation and erosion).

The authors of [5] designed a new decision based morpho filter for de-noising salt and pepper noise. The authors of [6] proposed an adaptive mathematical morphology for impulse noise. In this work, the authors emphasized on adjusting the size and shape of structuring element based on the local information of an image for de-nosing impulse noise. The authors of [7] designed a medical image enhancement using morphological transformation system to improve the quality of an image. Besides, this study utilized a mask of an arbitrary size and keeps changing its size until an optimum enhanced image is obtained from the transformation operation.

The systems described in reference [1], [2], [3], [5], [6] and [7] only emphasized how to remove noise by using the combination of two basic morphological operations and a mask of different shape and size. However, they only focus on improving the quality of image by de-noising an image. They did not mention about the reduction in computational time of conventional morphological and adaptive morphological approach.

3. Conventional Morphological Operations

Erosion and dilation are indeed the two most basic morphological operations for removing or attaching single pixels layer referring to structuring element. In this research, morphological operations are aimed at binary images with two possible pixel values, 0 and 1 [6]. Binary images can be described as follows:

$$I_B(x, y) \in \{0, 1\} \quad (1)$$

Where I denotes a binary image, and (x, y) is coordinate. Section 4.1, 4.2 and 4.3describes structuring element, dilation and erosion in details, respectively.

3.1. Structuring element

Besides, the structuring element is specified by a matrix that contains only the values 0 and 1. In other words, it is a small binary image. Thus, it can be a 3 ×3 square or 9 × 9 square image. Structuring element can be expressed as follows:

$$H\,(i, j) \in \{0, 1\} \quad (2)$$

Where H means structuring element and (i, j) is coordinate.

H=

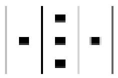


Figure1. 3 × 3 binary structuring element. 1 is marked with ■ and 0 cells are empty.

This research primarily uses the 3 × 3 structuring element as illustrated in Figure 1.

3.2. Dilation

Dilation grows or thickens an object in an image. As a set operation, it is defined as

$$I_B \oplus H \equiv \{(p + q) \mid \text{for all } p \in I,\ q \in H\} \quad (3)$$

The point set $I_B \oplus H$ produced by a dilation is the sum of all possible pairs of coordinate points from the original sets I_B and H.

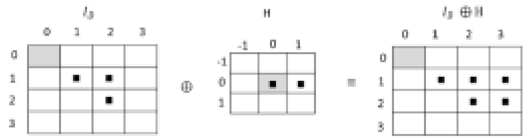


Figure 2. Simple Binary Dilation

$I_B \equiv \{(1, 1), (2, 1), (2, 2)\}$, $H \equiv \{(0, 0), (1, 0)\}$

$$\begin{aligned} I_B \oplus H \equiv \{&(1, 1) + (0, 0), (1, 1) + (1, 0),\\ &(2, 1) + (0, 0), (2, 1) + (1, 0),\\ &(2,2) + (0,0),\ (2,2) + (1,0)\} \end{aligned}$$

The image I_B is dilated with the structuring element H and $I_B \oplus$ H the result of dilation operation. The structuring element H is replicated at every foreground pixel of the original image, I_B.

3.3. Erosion

The quasi-inverse of dilation is the erosion operation, again defined in set notation as

$$I_B \ominus H \equiv \{p \in \mathbb{Z}^2 \mid (p + q) \in I, \text{ for all } q \in H\} \quad (4)$$

This operation can be expressed as follows. A position p is contained in the result $I_B \ominus$ H if (and only if) the structuring element H-when placed at this position p-is fully contained in the foreground pixels of the original image.

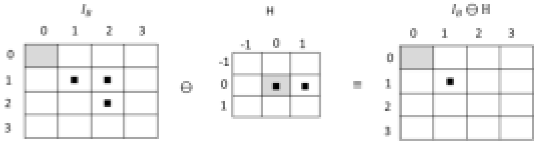


Figure 3. A simple example for binary erosion

$I_B \equiv \{ (1,1), (2,1), (2,2) \}$, $H \equiv \{ (0,0), (1,0) \}$

$I_B \ominus H \equiv \{ (1,1)\}$ because $(1,1) + (0,0) = (1,1) \in I_B$ and $(1,1) + (1,0) = (2,1) \in I_B$

As stated in conventional morphological operations, dilation and erosion are often used together in practice. Therefore, the methods that are erosion followed by dilation and dilation followed by erosion are widely used approaches for various image enhancements with different structuring elements.

4. Problems and Solutions

This paper mainly emphasizes on removing noise from rainfall radar images that is apparently a mixture of noises closely similar to salt and pepper noses as illustrated in Figure 4, the left hand-side image (original rainfall radar image). When we apply erosion with 3×3 structuring element to noisy rainfall image, it could perfectly remove out all noises from no rainfall area as shown in Figure 4, the right-hand side image.

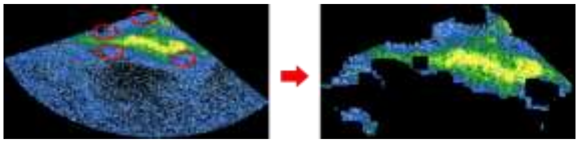


Figure 4. Noisy radar image and noise clean rainfall radar image

However, unfortunately, the erosion function operation left the image with big holes by wiping out the essential rainfall areas as illustrated in Figure 4, the right-hand side image, as those hole-areas occupy light rainfall areas having lower pixel density than heavy rainfall areas. In the rainfall radar images, the pixel density of heavy rainfall area is mostly thicker than light rainfall area as well as the pixel density of light rainfall area is thicker than no rainfall area.

To overcome the problem of leaving hole-areas, the method of erosion followed by dilation approach or dilation followed by erosion can accomplish the problems discussed above as well as remove all noise perfectly, as reported by Figure 5, the left-hand side image.

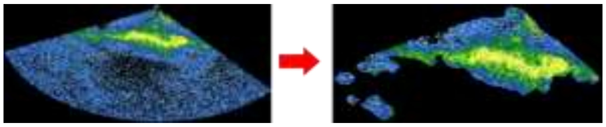


Figure 5. Noise cleaned rainfall image by the conventional approach

However, the repeatedly sequential function of erosion followed by dilation or dilation followed by erosion raises time complexity, the usage of memory, and gradually losing the essential rainfall pixel. Thus, the computational complexity of erosion and dilation for each pixel can be described in N + 2 operations, where N denotes the number of pixels in the structuring elements. Besides, this conventional method cannot completely fill back the holes with pixels as near as the original ones.

Let us first take a closer look at the essential areas, as shown in Figure 6. Figure 6 (a) and (b) have different problems. Figure 6 (a) shows the rainfall areas occupy the thicker pixel density. Those areas that are parts of rainfall region already have the proper enough pixel density. If we apply erosion operation to those areas, it will truly remove out the important pixels from those areas and leave them with holes. Therefore, protecting from eliminating the important pixels, the function of dilation is only enough to operate on those areas, instead of running two approaches on the same area directly.

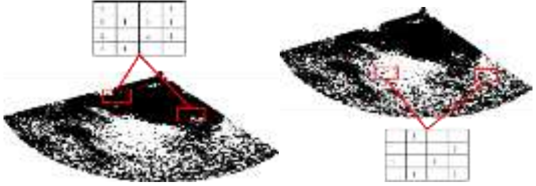


Figure 6. (a) and (b) Pixel density of a radar image

For the second case, the pixel density in the red rectangle regions, as illustrated in Figure 6 (b), is fairly low compared with the red rectangle regions in Figure 6 (a). Literally, the noises like salt and pepper are more likely to be naturally injected into those kinds of areas. Thus, for those areas, the function of erosion is more appropriate to take a role of removing noise, instead of assigning dilation to those area.

As discussed above, some regions more prefer erosion to dilation operation, while some regions are reasonable to use dilation operation with respect to structuring element. According to this analysis, an adaptive approach comes to us to handle those problem. The adaptive approach measures the pixel density of targeted region and examines which morphological operation (erosion or dilation) is perfect to take a role of removing noise. As the adaptive approach chooses properly one of two operations (dilation or erosion) according to the local pixel density, it intuitively solves the problem of time complexity, removing noise and protecting the essential rainfall area from wiping out during the repeatedly filtering noise process.

5. Adaptive morphological operation

This paper primarily focuses on upgrading computation time, filtering nose and protecting from wiping the important pixel area out during the noise filtering process. As discussed in the previous section, the adaptive morphological operation is a simple solution by selecting the appropriate one of two operations (erosion and dilation) respecting with the average pixel density of a targeted region in a binary image. Before selecting one of those two operations, it specifies the size of the interested filter region, R(x, y) as shown in Figure 7. The computation of the density of the filtered region is given by:

$$D(R) = \frac{1}{(x\times y)}\sum_{x=0}^{19}\sum_{y=0}^{19} R[x,y] \tag{5}$$

The size of the interested filtered region is also an important parameter of the filter because if the size is large, then the computation time is high, while if the size is small, the filtering noise may not work with a satisfied result. In the work of this paper, the size of the region is defined as 20× 20 square matrices

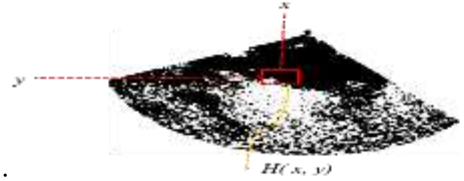


Figure 7. Filtered region R(x, y)

$$R'(x,y) = \begin{cases} R \oplus M \;, & if\ D(M) > threshold \\ R \ominus M \;, & otherwise \end{cases} \tag{6}$$

where $R'(x,y)$ is the new noise clean region after applying one of those two operations and in this research, R is the targeted region of a binary image, and the threshold is defined 10% according to our experimentations.

Algorithm 1 implements an adaptive morphological operation for a binary image. First, it converts radar image into the weighted image in order to avoid destruction of the pixels during transforming from a gray image into a binary image. The rainfall radar image has the intensity of rainfall levels represented by 15 different colors as shown in Figure 8. For the purpose of using weighted value, those colors are categorized into five groups, heavy raining, semi raining, fair raining, normal raining, and no raining. The heavy raining group includes red, pink and light pink. The next three colors, yellow-orange, yellow and light yellow are in semi raining group. The third group includes dark green, light green, and green. The fourth group has dark blue, blue and light blue and the rest colors are for the fifth group. Then, we specify the weighted value, 200, for heavy raining, 150 for semi raining, 100 for fair raining, and 70 for normal raining.

Dilation (*q*,*H*)

```
Input: a pixel of binary image, q.
a binary structuring image, H.
Output: $q'$, the dilated region = $q \oplus H$
$q' \leftarrow 0$
for all $(p) \in H$ do
    $q' \leftarrow q \oplus H$
Return $q'$
```

Erosion (*q*, *H*)

```
Input: a pixel of binary image, q.
a binary structuring image, H.
Output: $q'$, the eroded region = $q \ominus H$
$q' \leftarrow 0$
    for all $p \in H$ do
        $q' \leftarrow q \ominus H$
Return $q'$
```

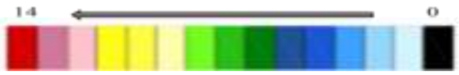


Figure 8. The intensity of rainfall levels

After converting into weighted value image, it is ready to transform into a grayscale image and then into a binary image. Then, according to the density of the region, it selects the appropriate operation.

Algorithm 1 : Adaptive Morphological Operation (*I*, *H*)

```
Input: Rainfall radar image, I, size of $M\times N$;
a binary structuring image, H.
Output: $I'$
$I_w \leftarrow I$ // convert image I into weighted image$I_w$
$I_G \leftarrow I_w$// convert image$I_w$ into gray image $I_G$
$I_M \leftarrow median\ (I_G)$//apply median filter to gray image
$I_B \leftarrow$ Binary($I_M$) // convert into binary image
Create map $I'$: $M\times N \to \{0, 1\}$
for all $(p) \in M\times N$ do
  for all q $\in M \times N$
    define R(x,y) in an image and compute the density
        every N
        $D(R) = \frac{1}{(20\times 20)}\sum_{x=m}^{x+20}\sum_{y=n}^{y+20} R[x,y]$
      if ($D(R)$> 10%) $q' \leftarrow$Dilation ( q, H)
          else Erosion$q' \leftarrow$ ( q, H)
      End
End
```

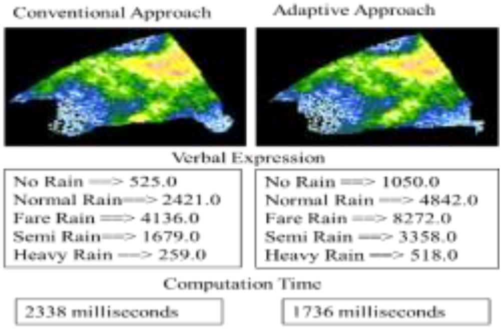
After that, the adaptive noise filtering approach starts to remove all noises from the image with high computational speed. At the final stage, the final noise clean binary image is converted into a color image by using a mapping approach.

6. Experimental results and analysis

In this section, we will discuss the experimentation of filtering noise from rainfall radar images. The results prove that the efficiency of the adaptive morphological operations by comparing with the conventional morphological approach based on 2011,2013, 2015 and 2016 radar images. In details, the total numbers of rainfall radar images are 116, 693 images (3.20 GB).

The performance of the adaptive approach is measured by using two factors: computation time, and the amount of important pixel value protected from wiping out as the accuracy. First, Figure 8(a) and (b), 10(a) and (b), 11(a) and (b), and 12(a) and (b) demonstrate the comparative study of conventional approach, adaptive approach, weighted value conventional approach, and weighted value adaptive approach using four radar images. As reported by those four images, the adaptive approach is more efficient than the conventional approach as expressed in each verbal expression and computation time. Likewise, it is obvious that the weighted value adaptive approach also achieves the accomplishment of higher computation time and accuracy than the weighted value conventional approach.

After discussing a comparative studying of the performance of weighted value adaptive approach using four rainfall radar images aiming at having closer look difference between among approaches, we did the experimentation again using 116, 693 (3.20 GB) radar images to prove that the performance of new approach with more confident results as demonstrated in Figure 13 and 14.



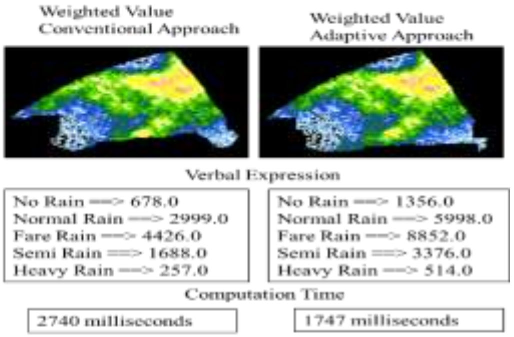
No Rain ==> 525.0
Normal Rain==> 2421.0
Fare Rain ==> 4136.0
Semi Rain==> 1679.0
Heavy Rain ==> 259.0

No Rain ==> 1050.0
Normal Rain ==> 4842.0
Fare Rain ==> 8272.0
Semi Rain ==> 3358.0
Heavy Rain ==> 518.0

2338 milliseconds

1736 milliseconds

Figure 9(a). A comparative study of conventional approach and adaptive approach



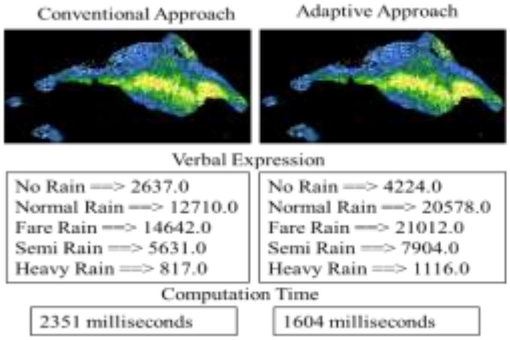
No Rain ==> 678.0
Normal Rain ==> 2999.0
Fare Rain ==> 4426.0
Semi Rain ==> 1688.0
Heavy Rain ==> 257.0

No Rain ==> 1356.0
Normal Rain ==> 5998.0
Fare Rain ==> 8852.0
Semi Rain ==> 3376.0
Heavy Rain ==> 514.0

2740 milliseconds

1747 milliseconds

Figure 9(b). A comparative study of weighted value conventional approach and weighted value adaptive approach



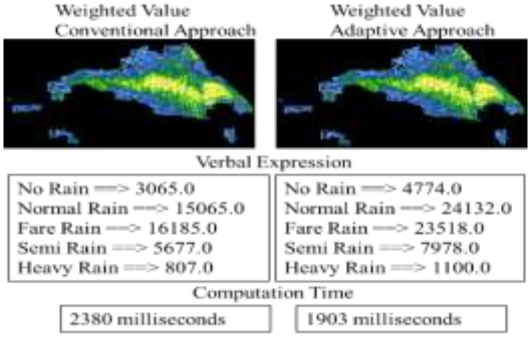
No Rain ==> 2637.0
Normal Rain ==> 12710.0
Fare Rain ==> 14642.0
Semi Rain ==> 5631.0
Heavy Rain ==> 817.0

No Rain ==> 4224.0
Normal Rain ==> 20578.0
Fare Rain ==> 21012.0
Semi Rain ==> 7904.0
Heavy Rain ==> 1116.0

2351 milliseconds

1604 milliseconds

Figure 10(a). A comparative study of conventional approach and adaptive approach



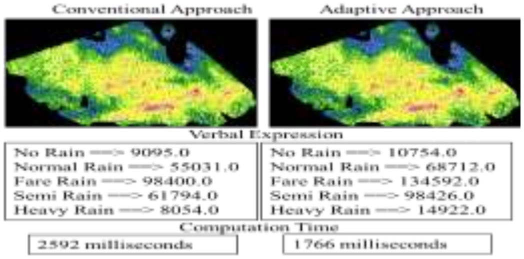
No Rain ==> 3065.0
Normal Rain ==> 15065.0
Fare Rain ==> 16185.0
Semi Rain ==> 5677.0
Heavy Rain ==> 807.0

No Rain ==> 4774.0
Normal Rain ==> 24132.0
Fare Rain ==> 23518.0
Semi Rain ==> 7978.0
Heavy Rain ==> 1100.0

2380 milliseconds

1903 milliseconds

Figure 10(b). A comparative study of weighted value conventional approach and weighted value adaptive approach



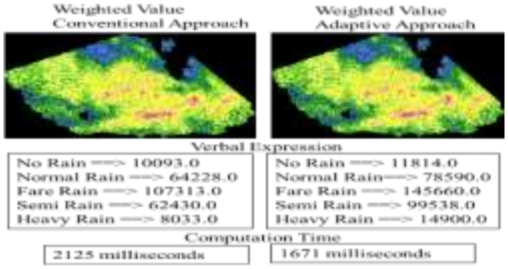
No Rain ==> 9095.0
Normal Rain ==> 55031.0
Fare Rain ==> 98400.0
Semi Rain ==> 61794.0
Heavy Rain ==> 8054.0

No Rain ==> 10754.0
Normal Rain ==> 68712.0
Fare Rain ==> 134592.0
Semi Rain ==> 98426.0
Heavy Rain ==> 14922.0

2592 milliseconds

1766 milliseconds

Figure 11(a). A comparative study of conventional approach and adaptive approach



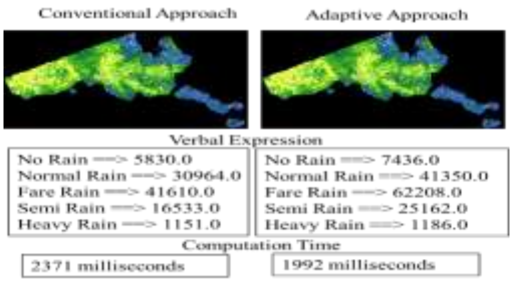
No Rain ==> 10093.0
Normal Rain ==> 64228.0
Fare Rain ==> 107313.0
Semi Rain ==> 62430.0
Heavy Rain ==> 8033.0

No Rain ==> 11814.0
Normal Rain==> 78590.0
Fare Rain ==> 145660.0
Semi Rain ==> 99538.0
Heavy Rain ==> 14900.0

2125 milliseconds

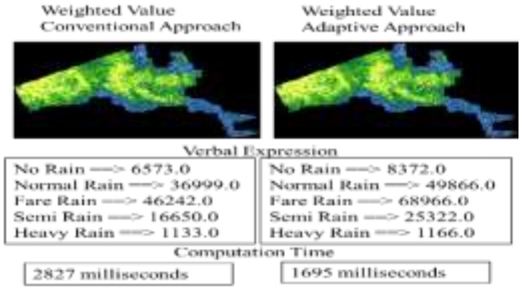
1671 milliseconds

Figure 11(b). A comparative study of weighted value conventional approach and weighted value adaptive approach



```
No Rain ==> 5830.0          No Rain ==> 7436.0
Normal Rain ==> 30964.0     Normal Rain ==> 41350.0
Fare Rain ==> 41610.0       Fare Rain ==> 62208.0
Semi Rain ==> 16533.0       Semi Rain ==> 25162.0
Heavy Rain ==> 1151.0       Heavy Rain ==> 1186.0
2371 milliseconds           1992 milliseconds
```

Figure 12(a). A comparative study of conventional approach and adaptive approach



```
No Rain ==> 6573.0          No Rain ==> 8372.0
Normal Rain ==> 36999.0     Normal Rain ==> 49866.0
Fare Rain ==> 46242.0       Fare Rain ==> 68966.0
Semi Rain ==> 16650.0       Semi Rain ==> 25322.0
Heavy Rain ==> 1133.0       Heavy Rain ==> 1166.0
2827 milliseconds           1695 milliseconds
```

Figure 12(b). A comparative study of weighted value conventional approach and weighted value adaptive approach

In Figures 13 and 14, the performance of two algorithms (adaptive and conventional approaches) is analyzed using 116, 693 images. As illustrated these figures, it is obvious that the adaptive noise filtering is more effective than the conventional approach because the adaptive approach filters the noise as perfect as the conventional approach. Furthermore, it maintains more pixels of essential rainfall areas than the conventional approach.

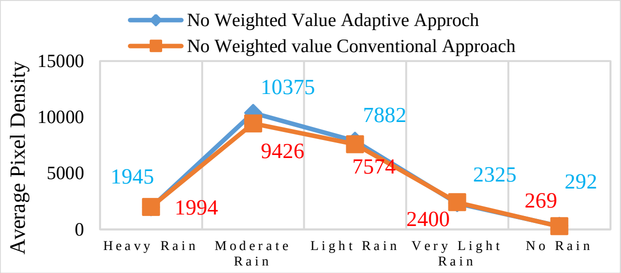


Figure 13. A comparative study of conventional approach and adaptive approach over four-year rainfall radar images

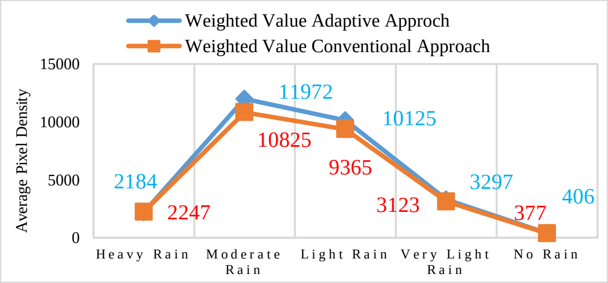


Figure14. A comparative study of weighted value conventional approach and weighted value adaptive approach over four- year rainfall radar images

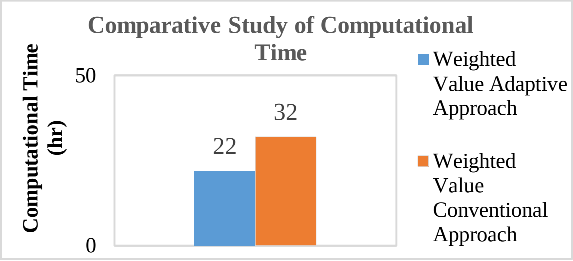


Figure 15. Computation time of weighted value adaptive approach and weighted value conventional approach over four-year rainfall data

As stated in Figure 15, the adaptive concept fairly boosts the computational time of repeated sequential noise filtering approach. The adaptive approach takes 22 hours for the noise filtering process using 3.20 GB radar images, while the conventional approach needs 32 hours. It reduces about 10 hours.

7. Computational Complexity

Suppose N × N is an image I, S × S is a structuring elements , H. The computation of conventional method, erosion I × S and dilation I × S of I by H would require $N^2 \times S^2$ for erosion and $N^2 \times S^2$ for dilation. Thus, the addition of two operations can be described as O $(2N^2S^2)$.

For the adaptation approach, each time it needs only one function (erosion or dilation) and the execution of density. Therefore, the computation of adaption approach can be specified as O(N^2S^2 + (N*N/20))→ O (N^2S^2 + N log N) . Because $2N^2S^2$ is always bigger than N^2S^2 + N*(N/20), we can conclude that O $(2N^2S^2)$>O (N^2S^2 + $N \log N$).

8. Conclusion

In this study, we propose the adaptive morphological approach aiming at upgrading the computational time and improving the noise filtering performance of conventional morphology. According to experimental results, weighted value adaptive approach accomplishes the task of filtering noise with 10 hours faster than the weighted value conventional approach as stated in Figure 15. As stated in Figure 13 and 14, adaptive approach maintains the important region 1.05 times (105%) more than the conventional approach, and similarly, weighted value adaptive approach protects the essential pixels 1.07

(107%) times more than weighted value conventional approach.

Finally, those results experimentally prove that the adaptive morphological approach is more efficient than the conventional approaches for noisy radar images. However, this adaptive approach considers only the pixel density of the targeted region in an image, not all the local information. For the future work, we will emphasize on improving the capability of noise filtering approach utilizing not only the density but also other local information of an image.

9. References

[1] V. Elamaran, H. N. Upadhyay, K. Narasimhan and J. J. Priestley, "A Case Study of Impulse Noise Reduction Using Morphological Image Processing with Structuring Elements", 2015, Vol. 8, No. 3, pp.291-303.

[2] N. Jamil, T. M. T. Sembok, Z. A. Bakar, "Noise Removal and Enhancement of Binary Images Using Morphological Operations", 2008 *International Symposium on Information Technology*, Kuala Lumpur, Malaysia, 26-28 August, 2008.

[3]A. Singhal, M. Singh, "Noise Removal and Enhancement of Binary Images Using Morphological Operations", *International Journal of Soft Computing and Engineering (IJSCE)*,November 2011, Volume-1, Issue-5, 2231-2307.

[4] W. Burger, M. J. Burge, "Digital Image Processing: An Algorithmic Introduction Using Java", *Springer-Verlag*, London, 2016.

[5] K.Priya and D. Pugazhenthi,"Salt and Pepper Noise Removal Algorithm by Novel Morpho Filter", *Journal of applied sciences*, Vol-14, No-9, pp 950-954, 2014.

[6] M.M.Javier,"Impulsive Noise Removal by Adaptive Mathematical Morphology", *Research in Computing Science*, Vol-112, No. 2016, pp. 65-76, May 25th 2016.

[7] R.Firoz et al. ,"Medical Image Enhancement Using Morphological Transformation", Journal of Data Analysis and Information Processing, Vol-4, pp. 1-12, January 28th 2016.

Natural Language Processing

Implementation of Recommender System Using Feature-Based Sentiment Analysis

Nyein Ei Ei Kyaw, Thinn Thinn Wai
University of Information Technology
Yangon, Myanmar
nyeineieikyaw@uit.edu.mm, thinnthinnwai@uit.edu.mm

Abstract

A recommender system aims to provide users with personalized online product or service recommendations to handle the increasing online information overload problem and improve customer relationship management. Collaborative Filtering (CF)-based recommendation technique helps people to make choices based on the opinions of other people who share similar interests. This technique has been suffering from the problems of data sparsity and cold start because of insufficient user ratings or absence of data about users or items. This can affect the accuracy of the recommendation system. User-generated reviews are a plentiful source of user opinions and interests. The proposed personalized recommendation model uses feature base sentiment analysis using ontology that extracts the semantically related features to find the users' individual preferences rather than rating scores in order to build user profiles that can be understood by user-based collaborative filtering recommendation model. The proposed model intends to alleviate data sparsity problem and to improve accuracy of recommender system by finding user preferences from review text.

Keywords- Collaborative Filtering (CF), Data sparsity, Review text

1. Introduction

The growth of the internet has made it much more difficult to effectively extract useful information from the available online information. We are suffering from information overload and being at a loss for the presence of too much information. A personalized recommendation system is one of the effective ways to solve this problem and has been used in many applications [6]. The mainstream of traditional recommendation approaches is usually based on the commonality among users i.e., similar users or entities are found by measuring the similarities of the common rating scores of users. However, the insufficiency of relevant data such as sparsity significantly weakens the effectiveness of these approaches due to the fact that there are often a limited number of common ratings among users.

User-generated online reviews have evolved into a pervasive part of e-commerce nowadays, as well as an essential focus of business intelligence and big data analytics. Both the online retail websites, like Amazon.com and Taobao.com, and the forum websites, such as Dianping.com and TripAdvisor.com are collecting tremendous amounts of online reviews.

Except for the ratings by users, the user reviews can offer much finer-grained information and have become a rich source to help detect the users' preferences. Most of the reviews contain users' opinions on various aspects of the target products/ services (referred to as entities). A user's preferences for the aspects of a certain entity are of great value in developing personalized recommendations [10].

Feature-based opinion mining is an attempt to identify the features of the opinion and classify the sentiments of the opinion for each of these features. The feature-based opinion mining of product reviews is a difficult task, owing to both the high semantic variability of the opinions expressed and the diversity of the characteristics and sub-characteristics that describe the products and the multitude of opinion words used to depict them [8].

The rest of the paper is organized as follows: Section 2 describes related works. Section 3 explains the background theory. Section 4 explains in details the architecture of the proposed system. Section 5 presents the performance evaluation of the proposed system. Section 6 describes the conclusion of paper.

2. Related Works

Several papers have addressed the problems to meet the personalized requirement of a user in various ways. Pallavi R. Desai, B. A. Tidke proposed a system that presents personalized recommendation lists and recommend the most appropriate items to the user by using weights of keywords are used to indicate user' preferences and a user-based collaborative filtering algorithm is adopted with Opennlp to generate appropriate recommendations [1]. Khushboo R. Shrote, Prof. A.V.

Deorankar proposed a system in which feedback analysis is done using sentiment analysis to recommend services. Keywords are used to indicate what the users prefer [2]. Susan Thomas, Jayalekshmi S proposed a system in which sentimental analysis on the reviews is done using Naïve Bayes, a machine learning technique to distinguish between the positive and negative reviews. It also uses MongoDB database to store the review detail [3]. Shakhy.P.S1, Swapna.H2 proposed a recommendation system which considers not only user reviews but also the temporal information about the location of the services. It uses Apache Mahout learning library and MongoDB to store reviews [4]. Dr. Kogilavani Shanmugavadivel and their colleagues proposed a system deals with the implementation of personalized rating to the services for hotel reservation system and booking of cars. This system performed opinion mining on the review at the sentence level using Bayes theorem and negation rule algorithm [5].

All the papers applied sentiment analysis on the reviews by using machine learning algorithms and then similarity of previous users and active user are computed and finally recommend the top N services to the new user. The candidate service sets (features of service) that system provide and domain thesaurus (similar terms associated candidate service) are manually specified. This is time-consuming and cannot relate the semantic meaning of the terms (features) more accurately.

3. Background Theory

3.1. Collaborative Filtering

Collaborative Filtering (CF) is considered the most popular and widely implemented technique in the recommender system. The underlying assumption of CF is that people with similar preferences will rate the same objects with similar ratings. Existing CF solutions can be categorized into two main classes: (i) memory-based and (ii) model-based methods. Memory-based solutions leverage similarities in users' behaviors and preferences to make inferences about missing values in the rating matrix. Memory-based algorithms (also known as Neighborhood-based) rely on the notion of similarity among users, or items, to predict the possible interest of a user on items that he/she has not seen (or rated) before.

Memory-based CF solutions are typically divided into two main categories: user-based and item-based. The user-based approach is based on the assumption that similar users typically rate the items in a similar way. Item-based CF focuses on the similarities among items. It is based on the assumption that similar items are rated in a similar way by the same user.

Model-based methods exploit the matrix values to learn a model, similarly to a classifier that trains a model from labeled data. The learned model is then used to predict the relevance of new items for the users. At present, collaborative filtering systems have a wide range of applications and provide customers with a good experience, but they still face a number of major problems such as data sparsity. When the data is very sparse, the accuracy of the recommendation from the collaborative filtering algorithm declines, which is a very big problem [9].

3.2. Sentiment Analysis

Sentiment analysis or Opinion mining is a part of Natural Language Processing which is used to analyze the opinions expressed by the different users. Sentiment analysis can be performed on three different levels namely at the document level, sentence level, and feature level.

Sentiment classification techniques can be divided into machine-learning approaches and dictionary-based approaches. However, despite the fact that machine learning approaches have made significant advances in sentiment classification, applying them to news comments requires the use of labeled training data sets. Dictionary-based approaches, on the other hand, can provide significant advantages, such as the fact that once they have been built, no training data are necessary [7].

3.3. Ontology

An ontology is a formal description of concepts in a domain of discourse (classes), properties of every concept describing various features and attributes of the concept, and restrictions on attributes. The concepts refer to various entities that may be any product or an organization. The use of ontology in feature-level opinion mining is to distinguish the domain related features by defining the classes in the domain and giving the relationships between the classes and instances [11].

4. Architecture of the Proposed System

The proposed system consists of three parts. They are

1. Ontology-based features identification.
2. Polarity identification using SentiWordNet.
3. User base collaborative filtering approach.

In the proposed system reviews text of users are firstly collected as the dataset. And then, preprocessing step is performed on the reviews text. After that, domain ontology is constructed to extract the features which have the semantic meaning from the reviews text.

In the second part, each of the specified features is classified into positive or negative polarity by using sentiment lexicons.

In the third part of the system, a recommendation process is performed. It has three sub-processes: user similarity computation, rating prediction on each item and ranking the items. Finally, the proposed system produces top N recommendation lists to the user.

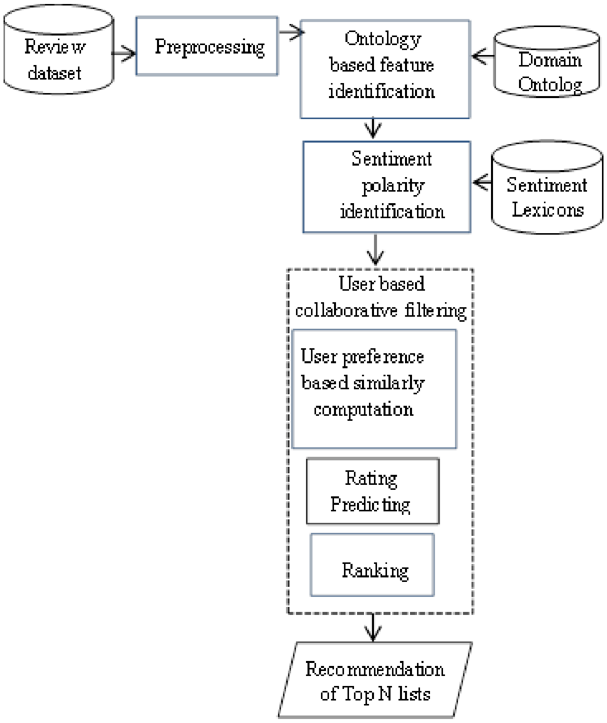

Review dataset
Preprocessing
Ontology based feature identification
Domain Ontolog
Sentiment polarity identification
Sentiment Lexicons
User based collaborative filtering
User preference based similarly computation
Rating Predicting
Ranking
Recommendation of Top N lists

Figure1. Proposed system architecture

4.1. Reviews Collection

In this phase, reviews about hotels from online website "https://www.kaggle.com/datafiniti/hotel-reviews" are collected as the dataset for the proposed recommendation system.

4.2. Preprocessing

After collecting the hotels' reviews from the respective website, preprocessing steps are carried out on these review sentences. For preprocessing, tokenization, stemming and stop words removing are carried out by using NLTK (Natural Language Toolkit).

Firstly, NLTK tokenizer tokenizes a sentence into words and punctuation. After that NLTK's Porter stemmer removes the commoner morphological and inflexional endings from words in English.

In the stop words removing process, stop words such as "a, an, the, that…etc" are removed. Hotel features in reviews are usually nouns or noun phrases, while user opinions are usually adjectives or verbs. POS tagging helps extracting such information from reviews. POS tagging is performed by using NLTK POS tagger to tag each word such as noun, adjective, verb, adverb, conjunction, preposition, and interjection. Sample tagging results from NLTK POS tagger is shown in Figure 2.

('hotel', 'NN'), ('wa', 'NN'), ('comfort', 'NN'), ('breakfast', 'NN'), ('wa', 'NN'), ('good', 'JJ'), ('-', ':'), ('quit', 'NN'), ('a', 'DT'), ('varieti', 'NN'), ('.', '.'), ('room', 'NN'), ('aircon', 'NN'), ('did', 'VBD'), ("n't", 'RB'), ('work', 'VB'), ('veri', 'RB'), ('well', 'RB'), ('.', '.'), ('take', 'VB'), ('mosquito', 'NN'), ('repel', 'NN'), ('!', '.'), ('realli', 'JJ'), ('love', 'JJ'), ('hotel', 'NN'), ('.', '.'), ('stay', 'VB'), ('on', 'IN'), ('the', 'DT'), ('veri', 'NN'), ('top', 'JJ'), ('floor', 'NN')

Figure 2. Sample tagging sets of hotel review

4.3. Domain Ontology Construction

Ontology is used to find the domain related features from the review sentences. Domain ontology is constructed by identifying concepts (classes), individuals, data type and object properties of the domain using the POS tagged words that resulted from the preprocessing step. Ontology construction is performed by using the Protégé 3.4 tool. Sample ontology of the hotel domain is shown in Figure 3.

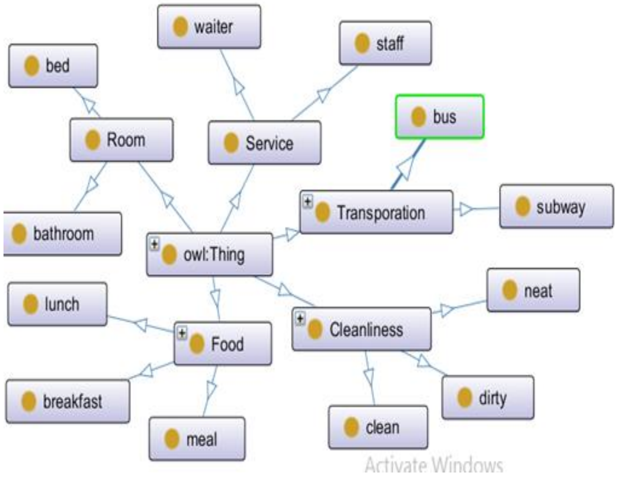


Figure 3. Sample ontology of hotel domain

4.4. Features Identification

In the features identification process, the features of user interests are extracted by using domain ontology. In this process, noun words that are identified by POS tagger is compared with the concepts of the domain ontology, and then these words are extracted as features.

4.5. Sentiment Word Extraction and Polarity Identification

In this process, opinion words such as adjective, adverb, and verb extracted from the review sentences are used to classify the positive or negative opinion of the users. Sentiment polarity identification process is carried out by using these opinion words and SentiWordNet 3.0.

4.6. Finding User Preferences

Based on the user's opinion, overall sentiment polarity score, the entity sets that the user reviewed and user opinions on each aspect are used to calculate users' preferences. This calculation is performed by the following equation [12].

$$User\ Preference\,(u_i, f_k) = \frac{\sum_{e_i \epsilon E_i} (S_{ij}\, S_{ijk})}{\sum_{e_i \epsilon E_i} S_{ij}^2 \sqrt{\sum_{e_i \epsilon E_i} S_{ijk}^2}}$$

Where,

S_{ijk} represents the opinions that user u_i comment on aspect f_k for entity e_i.

S_{ij} represents the overall sentiment polarity score that user u_i assign to entity e_i.

E_i is the entity set that user u_i reviewed.

4.7. Similarity Calculation to Predict Rating and Ranking Items

After getting the users' preferences, user based collaborative filtering algorithm which is based on user preferences is used to predict rating and ranking items.

Firstly user similarity calculation is carried out by using the preferences of the target user and other users. This calculation is done by using Cosine similarity method.

$$\mathrm{Sim}(u_i, u_m) = \frac{\sum_{k=1}^{|F|} (up_{ik} - \overline{up}_i)\ (up_{mk} - \overline{up}_m)}{\sqrt{\sum_{k=1}^{|F|} (up_{ik} - \overline{up}_i)^2 (up_{mk} - \overline{up}_m)^2}}$$

Where,

up_{ik} is the preference of target user u_i for aspect set F= (f1, f2,..., fk).

$\overline{up}_i$ is the preference of target user u_i for entity e_i.

up_{mk} is the preference of other user u_m for aspect set F= (f1, f2,..., fk).

$\overline{up}_m$ is the preference of other user u_m for entity e_i.

And then, items that have higher candidate score are ranked to the user. This score is calculated using the following equation based on the user preference with respect to other users with similar preference [12].

$$CS(u_i, e_j) = \overline{S_i} + \frac{\sum_{u_m \epsilon u_M} \mathrm{sim}(u_i, u_m) \cdot (S_{mk} - \bar{S}_m)}{\sum_{u_m \epsilon u_M} \mathrm{sim}(u_i, u_m)}$$

Where,

$\overline{S}_i$ represents the overall sentiment polarity scores of user on entity e_j.

u_M represents set of users who write reviews about entity e_j.

S_{mk} represents the opinion of similar user on aspect fk.

$\overline{S}_m$ represents the overall sentiment polarity scores of similar user on entity e_j.

5. Performance Evaluation

In the proposed system, performance evaluation is carried out by calculating precision and recall. Precision is the proportion of recommended items in the top N set that are relevant. Larger the precision better the recommendations. It is calculated by the following equations [14].

$$\text{Precision} = \frac{\text{Number of recommended items that are relevant}}{\text{Total number of recommended items}} \quad (1)$$

Recall is the proportion of relevant items found in the top N recommendations. It is calculated by the following equation.

$$\text{Recall} = \frac{\text{Number of recommended item that are relevant}}{\text{Total number of relevant items}} \quad (2)$$

The evaluation was based on the threshold value 3.5 which is specified by user and 200 rows from dataset for ten hotels. The actual rating greater than 3.5 regard as relevant items and user preference greater than 3.5 regard as recommended items. Number of recommended items that are relevant achieved from the intersection of the relevant items and recommended items.

Based on these assumptions, precision and recall are calculated by equations 1 and 2 respectively. Precision 75% of the recommended items was actually relevant to the user. Recall 70% of relevant items were recommended in the top N lists. The results are shown in Figure 4.

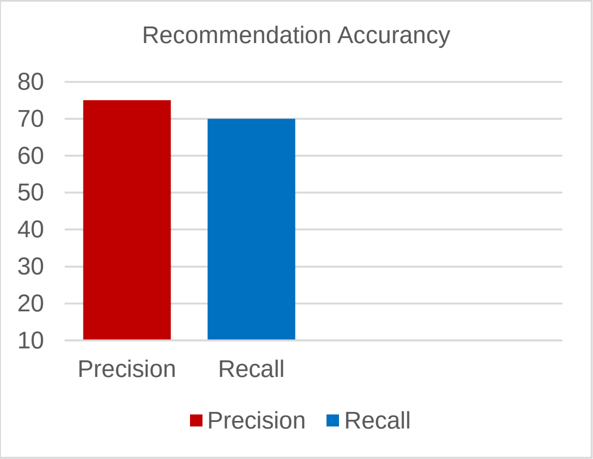


Figure 4. Performance evaluation of proposed system

6. Conclusion

In this paper, the proposed recommender system employs feature based sentiment analysis which use ontology at the feature extraction process to collect the semantic meaning of the features in review text. Then, customer preference and customer similarity are calculated based on feature words and sentiment polarity. Finally, the proposed system produces top N recommendation lists to the users.

7. References

[1] Pallavi R. Desai, B. A. Tidke, "A Survey on Smart Service Recommendation System by Applying Map Reduce Techniques", International Journal of Science and Research (IJSR) 2014.

[2] Khushboo R. Shrote, Prof. A.V. Deorankar, "Sentiment Analysis Based Feedback Analysed Service Recommendation method For Big Data Applications", International Journal of Scientific & Engineering Research 2016.

[3] Susan Thomas, Jayalekshmi S., "Recommendation System with Sentimental Analysis using Keyword Search", international journal for advance research in engineering and technology 2015.

[4] Shakhy.P.S1, Swapna.H2,"Improved Keyword Aware Service Recommendation System for Big Data Applications", International Journal of Innovative Research in Computer and Communication Engineering 2015.

[5] Dr. Kogilavani Shanmugavadivel, Dr. Thangarajan Ramasamy , Dr. Malliga Subramanian," Semantic Ranking Based Service Recommendation System using MapReduce on Big Datasets" , International Journal of Advances in Computer and Electronics Engineering 2017.

[6] Cheng Xiao, Dequan Zheng, Yuhang Yang, Automatic Domain-Ontology Structure and Example Acquisition from Semi-Structured Texts Sixth International Conference on Fuzzy Systems and Knowledge Discovery 2009.

[7] Anisha P Rodrigues, Dr. Niranjan N Chiplunkar,"Mining Online Product Reviews and Extracting Product features using Unsupervised method", 978-1-5090-3646-2/16/$31.00 ©2016 IEEE.

[8] Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M.Moreno, V., Fraga, A., Sánchez-Cervantes, J.L., Feature-Based Opinion mining through ontologies, Expert Systems with Applications (2014), doi: http://dx.doi.org/10.1016/j.eswa.2014.03.022.

[9] Mattia G. Campana, Franca Delmastro "Recommender Systems for Online and Mobile Social Networks: A survey", *IIT-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy, 2017 Elsevier B.V. All rights reserved.*

[10] Yue Ma, Guoqing Chen, Qiang Wei, "Finding users preferences from large-scale online reviews for personalized recommendation", Springer Science+Business Media New York 2016.

[11] Drashti Naik, Jitali Patel, "Feature Extraction from Product Review Using Ontology", International Journal of Computer Sciences and Engineering Vol.5 (8), Aug 2017, E-ISSN: 2347-2693.

[12] Nan Jing, Tao Jiang, Juan Du, Vijayan Sugumaran, "Personalized recommendation based on customer preference mining and sentiment assessment from a Chinese e-commerce website", Springer Science+Business Media, LLC 2017

[13] https://www.kaggle.com/datafiniti/hotel-reviews.

[14] https://medium.com/@m_n_malaeb.

Social Media Text Normalization

Thet Thet Zin
University of Computer Studies (Thaton)
ttzucsy@gmail.com

Abstract

Recent years some researchers interested in text normalization over social media, as the informal writing styles found in Twitter and other social media data. These informal texts often cause problems for Natural Language processing applications such as various mining research or translation on social media data. Today Facebook supports English translation of post and status for Myanmar Language. However, Most of the translation is not relevant for Myanmar words meaning. Complex nature of Myanmar language's syntactic structure, informal writing style, slang words and spelling mistakes are challenge in social media text translation work. This paper proposed text normalization that can be deployed as a preprocessing step for opinion mining, machine translation and various Natural Language Processing (NLP) applications to handle social media text. There are three steps in this work: Firstly, candidate words for normalization are selected from the collected raw dataset. In this case, Out-Of-Vocabulary (OOV) words are extracted for normalization. However, not all OOV words need to be normalized. Therefore, ill-formed words are detected from OOV words list for normalization. Second, slang words dictionary is generated for this work. Third, text similarity methods are applied to ill-formed words for normalization. Evaluation will be done on translation by applying normalization in pre-processing step. For translation, Myanmar-English machine translation [14] is used. The experimental results improve by applying proposed normalization to the translation work especially for social media text.

Keywords- informal text, social media, normalization, Out-Of-Vocabulary word (OOV), translation

1. Introduction

Now, nearly all people use user-oriented media such as social networking sites, blogs and micro blogging services. This led to a rapid increase in the need to understand casual written style, which often does not conform to rules of spelling, grammar and punctuation. Social media text is usually very noisy and contains a lot of typos, ad-hoc abbreviations, phonetic substitutions, customized abbreviations and slang language. The quality of text varies significantly ranging from high quality newswire-like text to meaningless strings. In Myanmar, mix usage of emotional voice and formal words change the meaning of the phrase. This is the big issue for translation especially for word-level translation work (eg. "ကောင်းတယ်ငိငိ"-"satire").This example cannot translate directly. Formal meaning of front word ("ကောင်းတယ်"-"good") is good. However, the meaning of whole word ("ကောင်းတယ်ငိငိ"-"satire") is satire to other. By combining the word ("ကောင်းတယ်"-"good") with the word ("ငိငိ"-informal word), become negative meaning of the word ("ကောင်းတယ်"-"good"). To handle this case, slang word or informal word dictionary is needed. Moreover, some write English pronunciation using Myanmar words: (eg. "တူဒေးမိနူး"-"today menu", wrong translated word is "nephew menu"). Therefore, social media text is often unsuitable as data for NLP tasks such as opinion mining, information retrieval and machine translation due to the irregularity of the language feature. Although average sentence length of social media text is small and generally commented on the posted text, it is not easier to find out the related context. Since social media text includes informal text, slang words, grammar and syntactic errors in the text. Moreover, some informal words have indirect meaning. To handle this case, informal or slang words dictionary and normalization process is needed to capture actual user's opinion for opinion mining.

In previous work [13] some way of preprocessing on the comments data is proposed to produce clean data. Aim in this paper is to normalize some ill-formed words such as multiword expression ("အောင်စေစေစေ"- "be successful") which cannot be solved in previous preprocessing work. Some normalization tasks are similarities with spell checking but differs in that ill-formed in text. Spell checker is also needed to normalize for machine translation. Nevertheless, ill-formed words or slang words like (eg. "ကွီး၊ မတ်မတ် is slang word. Formal words is "ကိုကြီး၊ မမ- brother,sister") tend to be considered beyond the scope of spell checking. In addition, the detection of informal words is difficult due to noisy context. The objective of this work is to detect ill-formed word and normalize to standard Myanmar word for translation. Similarity method is applied to OOV words. Category of OOV will discuss in the next section. In this approach a list of candidate word for normalization is generated firstly. Then slang word dictionary for Myanmar language is generated for normalization. Finally, similarity calculation is done between ill-formed words and candidate words. Proposed method supports to improve F-score and BLEU score in translation work.

Contributions in this paper are as follows: (1) studying the OOV word distribution of text and analyze different sources of non-standard orthography in data; (2) generating a slang words dictionary based on social media text; (3) detecting ill-formed words for normalization work exploits dictionary lookup and word similarity without requiring annotated data; (4) demonstrating the method better support for translation over social media text.

2. Related Work

Research aimed at the specific problem of normalizing casual Myanmar language is relatively rare. Some researcher fixed this problem by using NLP tools to social data. NLP tools for Myanmar language are rare in present time. The normalization approach is especially attractive as a pre-processing step for applications, which rely on key word match or word frequency statistics. For example, "အောင်စေ အောင်စေစေစေစေ အောင်စေစစစစ- "be successful") all attested in a Facebook comments corpus – have the standard form "အောင်စေ"- "be successful"; by normalizing these types to their standard form, better coverage can be achieved for keyword-based methods, and better word frequency estimates can be obtained.

The range of problems presented by user-generated content in online sources go beyond simple spelling correction; other problems include rapidly changing out-of-vocabulary slang, short-forms and acronyms, punctuation errors or omissions, phonetic spelling, misspelling for verbal effect and other intentional misspelling and recognition of out-of vocabulary named entities [2]. To discover the sequential dialogue structure of open-topic conversation in Twitter, [3] proposed unsupervised based conversation model. They compared Bayesian inference to Expectation-Maximization (EM) on conversation ordering task, showing a clear advantage of Bayesian methods. Hany and Arul [4] propose another unsupervised learning of the normalization equivalences from unlabeled text. They presented contextual graph random walks for social text normalization. Their proposed system based on constructing a lattice from possible normalization candidates and finding the best normalization sequence according to an n-gram language model using a Viterbi decoder. In addition, used random walks on a contextual similarity graph constructed form n-gram sequences on large unlabeled text corpus. They evaluated the approach on the normalization task as well as machine translation task. They figured out some limitations in normalization task and did not consider for mixed usage of words (eg, text include Myanmar and English words). Qi and other researchers [5] proposed Chinese-English mixed text normalization work. Experimental results on a manually annotated micro blog dataset demonstrate the effectiveness of their proposed method. From the results, this method can significantly benefit other NLP tasks in processing mixed usage of Chinese and English. Some researchers divide the text normalization problem into two sub-categories: word-based and character-based normalization. The word-based normalization turns non-standard words such as slang, acronyms and phonetic substantiation into standard dictionary words. Character-based normalization transforms the raw text through substituting the irregularly used characters with proper ones. Unsystematic usage of Latin alphabets (UULA) is presented by Osman and Ruket on noisy Uyghur text [6]. UULA normalization is character-level normalization. The noisy channel model and the neural encoder-decoder model are proposed and compared as normalizing methods. The noisy channel model views the problem as a spell-checking problem, while the neural encoder-decoder model views it as a machine translation problem. Both of them return highly accurate results on restoration and recommendation tasks on the synthetic dataset. However, their accuracy on real dataset would benefit from further improvement. To improve their performance on the real dataset, one possible strategy is to consider other noisy factors appearing in the real dataset.

Now especially at the social media in Myanmar, most users use informal writing style and appearing many slang words. Grammar and syntactic mistake also found in social media text. These cause issue for translation processes. Therefore, normalization for Myanmar social media text is needed. According to my knowledge, it is very rarely related research work in this area.

3. Data Analysis

As already described above, most of users, write status and comments using informal text in social media. This data need to be normalized for further processing. In this section, the dataset is examined for better understanding of the nature of data collected from Facebook. According to analysis, they use abbreviation (short form or acronym), slang word, mix typing usage (Myanmar and English word eg. (tmrသွားမှာလား- will tomorrow go?), multiword expressions, emotion icons and syntactic mistake. During the present time, many slang words in Myanmar language appeared via Facebook. Data from Facebook is collected by using Facebook API. Firstly, data is analysis into two parts formal and informal text. In the informal text category, abbreviation (short form and acronym), non-dictionary slang words, multiword expressions, mixed usage of two languages, orthographic mistakes, omission of vocabulary, combining two or three words to one slang word and further categories: Named entity, swear-word censor avoidance, emotion icon have been included. For this work, Facebook status data is extracted from 1st June 2018 to 1st July 2018. There are 20,897 sentences with length between 20 and 35. To analysis formal and informal text percentage in collected data, we selected 1,000 sentences from dataset randomly. 68% of selected sentence use formal writing style and 32% are using informal style. Most of the informal texts are phonetic substantiation into standard dictionary words. The detail analyze of informal text is described in the table.

Table1. Category of informal text

Category	Percent	Example
Abbreviation (short form or acronym)	5%	ဝကခ (ဝန်ကြီးချုပ်) ၊ မလမ (မော်လမြိုင်)
Omission of vocabulary	20%	ခေး(ကလေး) ၊ ကွီး(ကိုကြီး) ၊ မွီးဂါး (မုန့်ဟင်းခါး)
Mix typing usage	5%	Trmလာမယ် ၊ okလေ
Multiword expression	10%	အောင်မြင်ပါစေစေစေစေ ၊ အောင်မြင်ပါစေ!!!!
Emotion icons	10%	☺ :P
Orthographic mistakes	20%	ဆိုက်ထားတဲ့အပင်လေး
Myanglish	10%	kaung par pi
Slang word	10%	အယ်လယ် ၊ လန်းချက်
Others **(** named entity and Swearword Censor **)**	10%	အောင်မင်္ဂလာအဝေးပြေး၊ စောက်သုံးမကျလိုက်တာ၊ ငိငိ ၊ ခီခီ

Spelling checker handles orthographic mistakes in the text. Now, it is not possible to integrate Myanmar word spelling checker in the process. Myanglish (using English words for Myanmar words pronunciation: eg 'နေကောင်းလား-how are you?'- *nay kg lar?*) words are difficult for normalization because writing style of one different from another. Moreover, detecting and analyzing emotion icons will do separate research in the future. Other category includes named entity and Swearword Censor Avoidance. Therefore, these four categories are out of this paper.

4. Ill-formed Words Detection and Normalization

Detecting ill-formed words for normalization is a challenging problem especially in social text for many reasons. First, it is not straightforward to define the Out-of-Vocabulary (OOV) words. Traditionally, an OOV word is defined as a word that does not exist in the vocabulary of a given system. However, this definition is not adequate for the social media text, which has a very dynamic nature. Many words and named entities that do not exist in a given vocabulary should not be considered for normalization. Moreover, same OOV word may have much appropriate normalization depending on the context and on the domain. Therefore, analysis for words for normalization is difficult in social media text. In this paper, four steps are proposed for detecting candidate words for normalization.

First, blank space and punctuation are removing from n-gram words sequence. Myanmar sentences are segmented by using Myanmar syllabus segmenter developed by knowledge engineering major students of University of Information Technology (UIT). Present time, this segmentation tool cannot upload into the university website. Accuracy of this tool is reported in previous work [13] and the project book of this tool can get in UIT's FCS department. The longest matching n-gram is applied to segmented Myanmar sentence. Tri-gram is the best for this work.

Second, Myanmar-English bilingual corpus for machine translation is used for this work. Formal Myanmar-English corpus are created since previous work [14] on open domain. There are 61,824 Myanmar words and 56,263 English words. Every n-gram word is searched in the corpus. Some words do not have individual meaning but they have meaning by combining other surrounding words. Therefore, longest matching is used in this work.

Third, Myanmar words have many prefix, suffix, counting words and stop words. There are also remove from OOV words list. There are 603 prefix, suffix and counting words [13]. After doing above three steps, the remaining words may be OOV words or candidate words for normalization.

Fourth, two similarity methods and slang words dictionary are used to calculate similarity value in candidate words and ill-formed words. Firstly, slang words are extracted by using created slang words dictionary. After extracting slang words, normalization is applied to these words. Format of slang words dictionary is shown in table2. For example; the word 'မွီးဂါး' is normalized to 'မုန့်ဟင်းခါး (Myanmar traditional food)' using slang word dictionary. After searching slang words, two ways of similarity are calculated on remaining OOV words in the sentence. Words can be similar in two ways lexically or semantically.

Words are similar lexically if they have a similar character sequence. String-based n-gram similarity is used for lexical similarity based on the number of shared n-gram. *X* is the candidate words and *Y* is the canonical form in the dictionary. Similarity measure is calculated using the number of shared n-gram between two words. If value is greater than or equal to 0.6, it assumes the two words are similar lexically and normalized the candidate word with the canonical form. For example the candidate word ('အောင်စေစေစေစေ-be successful') is normalized to the canonical form (‘အောင်စေ’). If lexical similarity value of the word is less than 0.6, semantic similarity for this word is calculated.

Semantic similarity is calculated for the words which are not similar lexically in the sentence. Words are similar semantically if they have the same words, used in the same way, used in the same context and one is a type of another. In this case, surrounding words of the OOV should be considered. *X* is the candidate words and *Y* is the canonical form of the word. Probability $i_n(X, Y)$ is calculated by using surrounding words of the candidate words in the sentence. For example; in the sentence “လေတဟုန်းဟုန်းတိုက်တယ် - The wind roared.” the word “တဟုန်းဟုန်း-roared” is OOV words and lexically similar word are not contains in the dictionary. But the dictionary contains the word “ဝုန်းခနဲ”. Probability of $i_n(X, Y)$ is calculated by adding surrounding n-grams words combination probability for two sentences “လေ*တဟုန်းဟုန်း*တိုက်တယ်” and “လေ*ဝုန်းခနဲ*တိုက်တယ်” – “The wind roared” and the corpus. If the probability value is greater than 0.5, it assumes that the two sentences has nearly the same content. One of the issues of this calculation is that the probability value totally depends on corpus size and sentences in it. This word does not need to normalize for translation. But it can reduce OOV words for translation process. Another challenge in this similarity calculation is the words order in the sentence. Eg. “*တဟုန်းဟုန်း***လေ**တိုက် တယ်” and “**လေ***ဝုန်းခနဲ*တိုက်တယ်”. At the present time, one direction combination probability is done in this work. To analysis detection of ill-formed words from OOV, we selected 1,000 sentences randomly. Analysis result shows in figure 1.

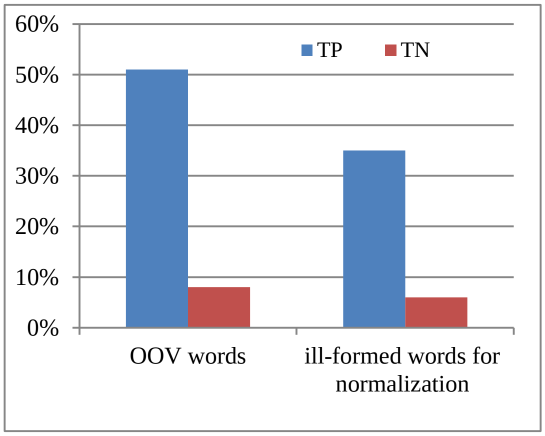


Figure1. Analysis of ill-formed words detection

According to analysis in figure 1, true positive is one that detects the condition (OOV words and ill-formed words) when the condition is present. True negative result is one that does not detect the condition when the condition is absent. 8% of true negative in OOV words include named entity and spelling error. Sometime, Myanmar words are similar semantically or word ordering (eg. အသုပ်စုံ၊အစုံသုပ်၊အသုပ် has the same translation word ‘salad’) but the system detects these words as OOV for domain. To analysis semantic similarity words, sometime it should consider the surrounding words also. 6% of true negative in ill-formed words include substitution phonic and slang words which do not include in dictionary (eg. ‘အကွက်တွေ’ is slang word. Formal word is ‘လှည့်ကွက်’. Translated word is ‘trick’). This is difficult to handle all slang words, which is used in social media. In the future, the dictionary will be perfect more than present time.

5. Normalization Lexicon Generation

We collected social media data from Facebook to generate normalization lexicon. Facebook statuses are collected for one month using Facebook API. There are 10,234 Myanmar sentences. Average words length of these sentences is 28. This paper uses a manually compiled and verified database, currently of a total of 805 entries. This amount is very small for normalization. These entries are either single words or phrases. At present time length of phrase entries are sets of two or three words. Each entry has been taken from separate sentences training data collected from Facebook status and comments. Database entries comprise of three columns: “the casual Myanmar word”, “regular word” (the corresponding dictionary Myanmar word) and “category”. One standard word has many relevant slang words. Database construction is an ongoing project, and intends to improve its coverage and quality further. Later, we will use unsupervised approach for generation for lexicon instead of manually compiled. Format of slang word dictionary is shown in table2.

Table 2 .Format of Myanmar slang dictionary

Casual word	Regular word	Category
ဝကခ၊ မလမ	ဝန်ကြီးချုပ်(prime minister)၊ မော်လမြိုင်(mawlamyine)	Abbreviation
မို့ဂါး	မုန့်ဟင်းခါး (Myanmar traditional food)	Omission of vocabulary
Today မီးနူး	တူဒေးမီးနူး(Today menu)	Mix typing usage
လန်းချက်	လှသည် (beautiful) မိုက်သည် (cool)	slang word

6. Experiential Result

We constructed a test set of 1,000 sentences with average sentences length 10 words are collected from social media, which are separated from training data set.

Furthermore, a test set is developed for evaluating the effect of the normalization process when used as a preprocessing step for translation work. A test set for human evaluation and BLEU scores. Human evaluation results are shown in figure 2. For translation, we used Myanmar-English translation proposed in [14]. We prepared translation reference for test set under the guideline of English lecturer. Precision and recall of words' translation are calculated by the following way.

$$precision \qquad \frac{correct\_words}{output\_lenght}$$

$$recall \qquad \frac{correct\_words}{reference\_lenght}$$

$$F-measure \qquad \frac{precision * recall}{(precision + recall)/2}$$

Table3. Example of translated sentence

Myanmar sentence	ဂရုတစိုက်နားထောင်ပါမတ်မတ်
Reference:	listen carefully sister
Translation without normalization	listen carefully *march march*
Translation with normalization	listen carefully sister

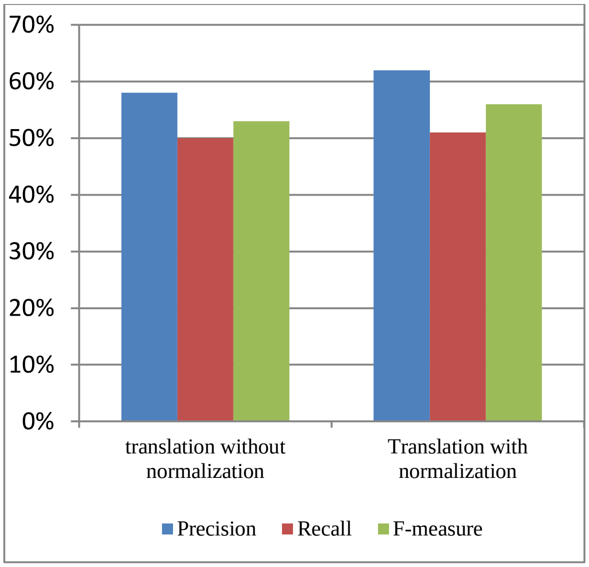


Figure2. Evaluation results for normalization

Most of the sentences can be translated in many acceptable forms. Thus, more than one reference sentences should be considered. One reference is considering for the results and ignores word order in the translated sentence. BLEU is a score for comparing a candidate translation of text to one or more reference translations. Higher numbers correspond to better translations. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. Some translated output is much too short, thus boosting precision, and BLEU doesn't have recall. We evaluate the translation based on 4-grams BLEU scores evaluation. The results are shown in the following table.

Table4. 4-gram BLEU score

	BLEU
without normalization	0.296
with normalization	0.373

Analysis on 1,000 sentences, 52% of precision for these sentences comes in human evaluation. This meaning that false positive rate increases in testing dataset. Ill-formed words detection has some errors. Recall on this case is higher than precision. This means that training data for this test set is reasonable enough. In the translation work, precision is higher than recall. Increasing the amount of training data will affect to the performance positively especially the recall. We also test translation without applying normalization process. The results show that translation uses normalization as a preprocessing step for a machine translation, which improved the translation quality by 3% in F-score.

In BLEU score of the translation with normalization, about 75.4% of the overall precision score comes from the uni-grams. 17.5% comes from the bi-grams; 4-grams contribute only 1.3%. The number of longer n-gram matches is smaller compared with shorter n-gram matches. we assume that the human evaluation scores are the most valid then the automatic metrics for only these 1,000 test set.

7. Discussion and Future Work

We first manually analyses the errors in normalization over the test set. There are two categories in error analysis. First is an error in ill-formed words detection. The most frequent in this category is caused by morphological variations, including (1) negations: In Myanmar language, negation is difficult to recognize for further processing because Myanmar has many negation forms. (eg. မရဘူးဟာ၊ ရမယ်လို့မထင်တာ၊ ရကိုမရတာ၊ မသိဘူးမရ ဘူး-all are negation form). It fails to normalize 32% of the negation words in the test set (2) syntactically or semantically ambiguity between words: about 23.5% of the words have ambiguous meaning (3) spelling errors: about: over 30% of the words have spelling errors (4) over 10%

of the words are missing in created slang words dictionary. The second category is false positive rate in OOV words. Some words have same meaning with different forms. This cause occurs in OOV words or ill-formed for normalization. This reduces precision of the translation. We already mention above that the translation uses normalization as a preprocessing step for a machine translation which improved the translation quality by 3% in F-score. Most errors in this case are that in social media text, some words cannot be translated directly. It cannot be translate by only considering surrounding words and sentence structure. For example: အထောင်းမှန်သမျှ-'all pounded food: such as pounded papaya' this word cannot translate English word "right" even through ("မှန်" is "right"). Some errors found in changing slang word to standard word for translation. Because some slang word has different meaning depend on content of the text. For example: (အကွက်တွေမိုက်တယ်- 'nice trick' ၊ သန့်နေတာဘဲ- 'neat and tidy'). Normalization process does not know these words need to normalize for translation work. Some output show that normalization process has done on words but normalized standard word is wrong for translation. To overcome this problem, consideration on content and sentence structure of the text include both normalization and translation processes. Moreover, slang word dictionary will need to powerful than before.

8. Conclusion

In this paper, normalization on a social media text is proposed that can be deployed as a preprocessor for MT to handle social media text. We analyzed the collected data and identified ill-formed words for normalization. Most informal text in social media based on spelling mistake, slang words and substation of phonic. Proposed informal text detection method shows accepted results. However, other experiment and methodology are needed to improve ill-formed word detection. Moreover, slang words database generation is an ongoing project. For effective normalization on social media text, powerful annotation corpus or effective unsupervised method is needed. Some limitations in proposed approach are found by analyzing output results: example mix type usage cause the problem for normalization. As an extension to this work, we will extend the approach to handle named entity and spelling mistake by integration Myanmar named entity recognition and spelling checker to the normalization on social text. Furthermore, the approach can be extended to handle semantically similar words problems for normalization. We hope the best results will outcome in the future.

9. References

[1] E.Clark and K.Araki, "Text normalization in social media: progress, problems and applications for a pre-processing system of casual English", *Pacific Association for Computational Linguistics (PACLING 2011)*, Published by Elsevier Ltd. 2011, pp. 1-10.

[2] E.Clark, T.Roberts and K.Araki, "Towards a pre-processing system for casual English annotated with Linguistic and Cultural Information", Proceeding of the fifth IASTED International Conference Computational Intelligence (CI 2010), August 23-25, 2010 Maui, Hawaii,USA.

[3] A. Ritter, C. Cherry and B. Dolan, "Unsupervised Modeling of Twitter Conversations", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, June 2010, Los Angeles, California, pp 172-180.

[4]H. Hassan and A.Menezes, "Social Text Normalization using Contextual Graph Random Walks", Proceedings of the 51st annual meeting of the association for computational linguistics, August 4-9 2013, Sofia, Blugaria, pp 1577-1586.

[5] Q.Zhang, H.Chen and X.Huag, "Chinese-English Mixed Text Normalization", WSDM , February 24-28 2014, New York, USA, Copyright 2014 ACM 978-1-4503-2351-2/14/02, pp 433-42.

[6] O. Tursun and R. Cakici, "Noisy Uyghur Text Normalization", Proceedings of the 3rd Workshop on Noisy User-generated Text, September 7, 2017, Copenhagen, Denmark, pp- 85-93.

[7] T. Baldwin and Y.Li, "An In-depth Analysis of the Effect of Text Normalization in Social Media", Human Language Technologies: The 2015 annual conference of the North American chapter of the ACL, May 31- June 5, 2015, Denver, Colorado, pp 420-429.

[8]B. Han and T. Baldwin, "Lexical Normalization of Short Text Messages: Makn Sens a #twitter ", proceedings of the 49th annual meeting of the association for computational linguistics, June 19-24, 2011, Portland, Oregon, pp 365-378.

[9] C. Henriquez Q and Adolfo Hernandez H, "A Ngram-based Statistical Machine Translation Approach for Text Normalization on Chat-speak Style Communications", CAW 2.0, April 21 2009, Madri, Spain.

[10] S. Rohatgi and M. Zare, "DeepNorm- A Deep learning approach to Text Normalization", ACM ISBN 123-4567-

24-567/08/06, IST597-003 Fall' 17, December 2017, State College, PA, USA.

[11] B. Han, P.Cook and T. Baldwin, “Automatically Constructing a Normalization Dictionary for Microblogs”, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, July 12-14 2012, Jeju Island, Korea, pp 421-432.

[12] S. Goyal and Er. Bedi, “SMS Text Normalization Using Hybrid Approach”, International Journal of Computer Trends and Technology (IJCTT), Volume 21 Number 2, ISSN: 2231-2803, march 2015, pp126-129.

[13] Zin et al.,“Domain-Specific Sentiment Lexicon for Classification”, the First International Conference on Advanced Information Technologies (ICAIT), November 1-2, 2017, Yangon, Myanmar.

[14] T.T.Zin, K.M.Soe and N.L. Thein, “Translation Model of Myanamr Phrases for Statistical Machine Translation”, the 2011 seventh international conference on intelligent computing, august 11-14 2011, Zhengzhou, Henan, China, copyright Springer-Verlag Berlin Heidelberg 2011, pp. 235-242.

[15]http://www.ucsy.edu.mm/GoNLP.do

Sentiment Aware Word Embedding Approach for Sentiment Analysis

Win Lei Kay Khine, Nyein Thwet Thwet Aung
University of Information Technology
Yangon, Myanmar
winleikkhine@uit.edu.mm, nyeinthwet@uit.edu.mm

Abstract

Nowadays, many business owners want to know the feedback of their products. If they get the feedback from customers, they can promote the quality of their products. So, Sentiment analysis has become a popular research problem to tackle in NLP field. It is the process of identifying whether the opinion or reviews expressed in a piece of work is positive, negative or neutral. We can apply sentiment analysis in brand monitoring, customer service, market research and analysis. Word embedding step is a problem in sentiment analysis of neural network models. Most existing algorithms for continuous word representation typically only model the syntactic context of words but ignore the sentiment of text. It is a problematic for sentiment analysis as they usually map words with similar syntactic context but ignore opposite sentiment polarity, such as good and bad, like and dislike. We solve this issue by proposing a method, sentiment-aware word embedding (SAWE). SAWE encodes sentiment information in the continuous representation of words by using (1) prediction the model and (2) ranking model. Finally, we evaluate our proposed method on IMDB movie review and twitter datasets, after that we prove our method outperform than other word embedding methods like word2vec and GloVe.

Keywords- Sentiment analysis, Natural Language Processing, Word Embedding, SAWE, Recurrent Neural Networks

1. Introduction

Nowadays, many business organizations want to promote their products in order to be successful. So, they survey about their products and use marketing strategies. It is expensive and time-consuming. If they use the sentiment application of their products, the above problem can be solved. If we do sentiment analysis, we first pass the data to the word embedding step.

Word embedding is a popular method for natural language processing (NLP) that aims to learn low-dimensional vector representations of words from documents. Due to its ability to capture syntactic and semantic word relationships, word embedding algorithms such as Skip-gram, CBOW and GloVe have been proven to facilitate various NLP tasks, such as word analogy, parsing, POS tagging, aspect extraction, etc... The majority of existing word embedding algorithms merely takes into account statistical information from documents. The representations learnt by such algorithms are very general and can be applied to various tasks. So, we propose sentiment-aware word embedding approach for sentiment classification task because it capture syntactic, semantic as well as sentiment information, unlike normal word embedding (word2vec and GloVe), which only capture syntactic and semantic information.

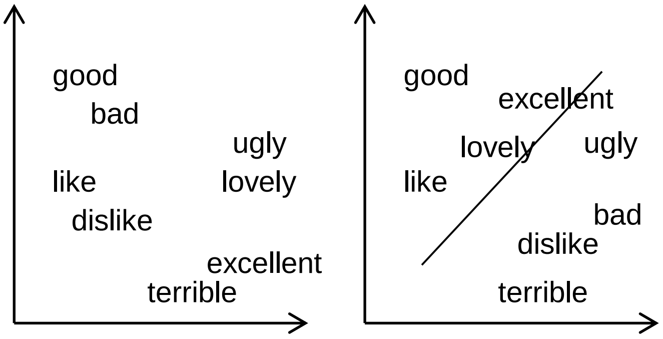


Figure 1. Normal word embedding (left) and sentiment-aware word embedding (right)

On the left of the figure(1) is normal word embedding which capture the syntactic context of words and the right one is capture and determine the sentiment of each word. So the positive and negative words are occupied separately in the vector space according to the polarity of the word. For example, the words "like" and "dislike" can appear in the same or similar context such as I *like* reading books or I *dislike* reading books. By merely looking at word co-occurrences, we would learn similar vector representations of "like" and "dislike" as these have similar lexical behavior. From a sentiment point of view, however, such vector representation should be very different as they convey opposite polarity. Hence, by incorporating prior sentiment knowledge about these two words, we can build more sentiment-aware word embedding and, hence, learn better distributional representations for sentiment analysis.

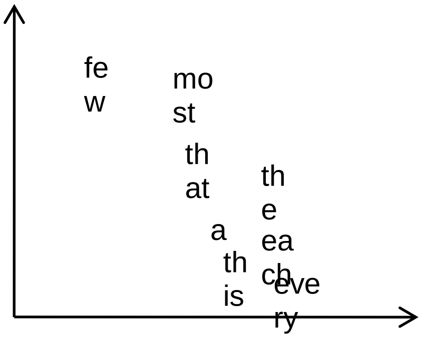


Figure 2. Word embedding for English determiners

The remaining paper is organized as follows. Section 2 describes the related works. Section 3 describes a methodology that is needed to implement sentiment classification. In Section 4, we present proposed system architecture and in section 5, we explain about datasets and experiment on these datasets. Finally, we conclude the paper in section 6.

2. Related Work

Word embedding is given to any method which converts words into numbers. We cannot directly feed text data to our machine learning or deep learning models. They won't work on strings of plain text. So, a natural language modeling technique like word embedding is used to map words to a corresponding vector of real numbers. Mikolov et al.(2013) introduce Continuous Bag-of-Words (CBOW) and Continuous Skip-gram, and release the popular word2vec toolkit. CBOW model predicts the current word based on the embeddings of its context words, and Skip-gram model predicts surrounding words given the embeddings of current word. [5].

Deep learning is the part of machine learning process which refers to Deep Neural Network. Neural Network is influenced by human brain and it contains several neurons that make an impressive network. It can have various numbers of nodes per layer, various numbers of hidden layers and weights connected between. The more layers a neural network has, the more complex model the network can learn. A neural network with multiple hidden layers is called Deep Learning [2]. Deep learning networks are capable for providing training to both supervised and unsupervised categories. It has been extensively applied in artificial intelligence field, computer vision, semantic parsing, and natural language processing many more [1].

In study by [3], the researchers proposed a novel Recursive Neural Deep Model (RNDM) to predict sentiment label based on recursive deep learning. In order to address the problem of little investigation on Chinese Sentiment analysis, they introduced a Chinese Sentiment Treebank and a powerful recursive deep model that can accurately predict the sentiment label on sentence level on movie review from social networks. The movie reviews were collected from http://movie.douban.com/. Finally, they reported that their RNDM obtains an accuracy of 90.8%, compared to NB (78.65%), ME (87.46%), and SVM (84.9%), so their RNDM achieves the highest accuracy in predicting binary sentiment label of sentence level.

The authors [4] proposed sentiment classification conducted on Japanese corpora. They trained and tested their sentiment classifier model, BiLSTM with Rakuten Merchant Review data. Their proposed can run without any dictionaries or features.

Another type of deep learning technique is Convolutional Neural Network (CNN). CNN consists of many layers that perform different functions. But CNN gives outstanding results in image processing and speech applications than text classification. The main problems of CNN are high computational cost and they need a lot of training data. And if you don't have a good GPU they are quite slow to train. And, if it is not having a good GPU, there will face a problem in training phase.

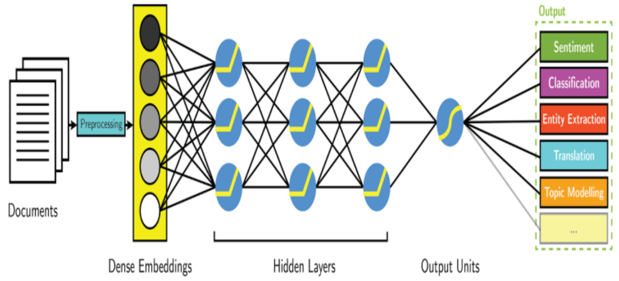


Figure 3.General deep learning based NLP

Figure 3shows the deep learning based NLP than any other classical NLP method. In our approach, we work with recurrent neural networks because it can handle more complex ways of connecting layers.

3. Methodology

In this section, we will discuss about word representations and popular techniques used in word embedding such as word2vec and GloVe.

3.1. Word Representations

Continuous word representation is commonly called word embedding, which attempt to represent each word as a continuous, low-dimensional and real-valued vector [6]. Word representation aims to represent aspects of word meaning. A straightforward way is to encode $w_i$as a one-hot vector, whose length is a vocabulary size by 1 in the w_i^{th} position and zeros everywhere else. However, such one-hot vector word representation only encodes the indices of words in a vocabulary.

3.1.1 Word2Vec. Word2vec is not a single algorithm, but a combination of two techniques – CBOW (Continuous bag of words) and Skip-gram model. Both of these are shallow neural networks, which map word in the target variable which is also a word. Both of these techniques learn weights which act as word vector representations.

CBOW is learning to predict the word by the context. Or maximize the probability of the target word by looking at the context. And this happens to be a problem for rare words. For example, given the context *yesterday was really [...] day* CBOW model will tell you that most probably the word is *beautiful* or *nice*. Words like delightful will get much less attention of the model, because it is designed to predict the most probable word. This word will be smoothed over a lot of examples with more frequent words.

On the other hand, the skip-gram is designed to predict the context. Given the word delightful it must understand it and tell us, that there is huge probability, the context is yesterday was really [...] day, or some other relevant context. With **skip-gram** the word delightful will not try to compete with word beautiful but instead, delightful+context pairs will be treated as new observations.

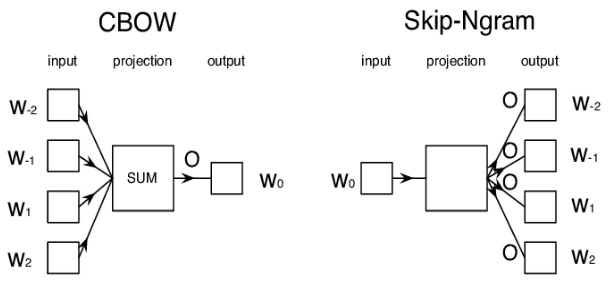


Figure 4.Continuous BOW a and Skip-Ngram approach used in word2vec

3.1.2 GloVe. An alternative approach for word embedding is called GloVe (Global Vectors) because the global corpus statistics are captured directly by the model. It is a model for distributed word representation. The model is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. It is developed as an open-source project at Stanford.

3.1.3 Sentiment Aware Word Embedding. The problem statement using word2vec and GloVe for word embedding is that they can't capture the sentiment information and they can only capture the syntactic and semantic information. So, the important feature for our work is to use Sentiment Aware Word Embedding (SAWE). SAWE can capture syntactic, semantic as well as sentiment information.

To implement word embedding in our system, we first describe standard context-based neural network methods for learning word embedding. Afterwards, we introduce our extension for capturing sentiment polarity of sentences which encode both sentiment and context level information. We then describe the integration of sentence level information for embedding learning. The discussion of the detail implementation of SAWE will explain in section 4.2.We implement our work with neural network layers, including *lookup*, *hTanh*, *linear* and *softmax*. For each neural layer, O_{layer}means the output vector.

Here, the proposed system builds sentiment-aware word embedding model in TensorFlow, an open source software library for high performance numerical computation. To work with TensorFlow, the Windows must be 64 bits-based Windows. We train our data on Window 7 64 bits, Core(TM) i7-4470, RAM 4.00 GB and require Python version 2.7 and above. We train and test the data on jupyter notebook in Anaconda Navigator.

4. Proposed System Architecture

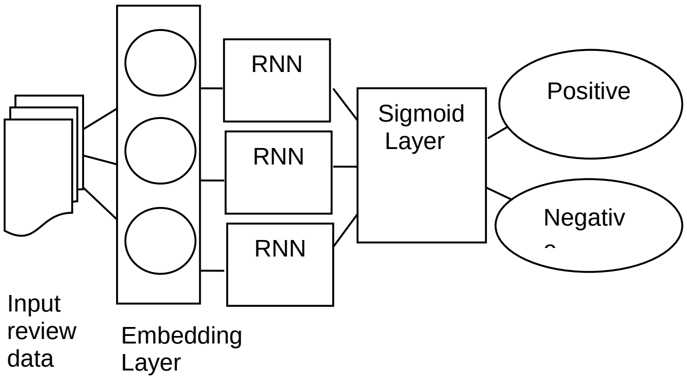


Figure 5.Overview of sentiment analysis using deep learning

According to the Figure 5, we firstly collected reviews texts data (Imdband Twitterdataset). Twitter data http://twitter.com by using Twitter API. Secondly, we need to pass words to an embedding layer and train up with SAWE model. Word embedding is not itself a deep learning technique, but it can turn raw text into a numerical form that deep nets can understand. For the word embedding, we used Keras in Embedding layer.

Keras is an open source neural network library. We run Keras on Tensorflow. We encode all words into word vectors. And then, we train Neural Network models on word vectors for classification. From the embedding layer, the new representations will be passed to Recurrent Neural Network (RNN).Finally, it goes to a sigmoid output layer. The sigmoid function takes any range real number and returns the output value which falls in the range of 0 to 1. Although there are many activation function, but we use sigmoid because it can predict if the text has positive or negative value for sentiment analysis. This is overview of my proposed system. But, my contribution is on sentiment-aware word embedding method on word embedding step.

4.1 Modeling Contexts of words

There are two parts in our proposed system. The first one is to model the context words and the second one is to model for sentiment polarity of the texts. In both of these sections, we will have to develop the prediction model and ranking model.

4.1.1 Prediction Model. In order to predict the contexts of words, we need to encode these contexts of words into word representation. This is called context prediction. Context prediction aims to predict the target word w_i based on its context words h_i. We need to predict surrounding words $h_i = \{w_{i-c}, w_{i-c+1}, \dots w_{i-1}, w_{i+1}, \dots w_{i+c-1}, w_{i+c}\}$. We build a prediction model analogous to the representative "context prediction" neural language model given by Bengio et al. [8]. They model the conditional probability $P(w_i| h_i)$ of predicting a target word $w_i$based on its contexts $h_i$by taking word embeddings as a parameter. The scoring function is a feed-forward neural network consisting of *lookup → linear→ hTanh→ linear→ softmax.*

Lookup layer also referred to as projection layer, which contains a lookup table LT $\in$ R $^{d\times|V|}$ which maps each word to its continuously vector.

d= dimension of each word

V= vocabulary size

LT= lookup table

The lookup operation can be viewed as a projection function that uses a binary vector idx_i, which is zero in all positions except at the i^th index.

$$e_i = LT \,.\, idx_i \in R^{1*d} \tag{1}$$

After that, we concatenate the embedding of context words as the output of lookup layer.

$$O_{lookup} = [e_{i-c}; \dots e_{i-1}, e_{i+1} \dots ;e_{i+c}] \in R^{1*d.2} \tag{2}$$

And then, the output of lookup layer is fed to a linear layer for dimension transformation is

$$O_{l1} = W_{l1} \cdot O_{lookup} + b_{l1,}\text{where} \tag{3}$$

W_{l1}= position-dependent weight

b_{l1}= bias of linear layer

O_{l1}= output vector of linear layer

In order to predict the probability of positive/negative polarity, we use *hTanh* (hard hyperbolic tangent) for its computational efficiency and effective in literature [9]. The output vector of hTanh is $O_{hTanh \in R}{}^{1*len}$

$$hTanh(x) = \begin{cases} -1 \; if \; x < -1 \\ x \; if -1 \leq x \leq 1 \\ 1 \; if \; x > 1 \end{cases} \tag{4}$$

The output layer is a softmax layer whose output length is vocabulary size. The probability of given sample from data D is defined as,

$$P(D|w,\theta) = \frac{\exp\left(f_\theta(w_{i,} \;\; h_i)\right)}{\exp\left(f_\theta(w_i \;,h_i)\right) + k.\exp\left(f_\theta(w^n \;,h_i)\right)} (5)$$

The score function $f_\theta(w,h)$quantifies the compatibility between context $h_i$and target word w_i, which can be naturally defined as a feed forward neural network consisting of *lookup → linear → hTanh→ linear*. The input of lookup layer is the concatenation of the current word *w* and context words *h*. The output is a linear layer with output length as 1, which stands for the compatibility between context *h* and word *w*. We implement $P(D|w,\theta)$ with a *softmax*layer and maximize the log probability of the *softmax*for parameter estimation

Finally, the prediction for context of word is

$$loss_{cPred} = \sum_{w \in T} log \; P(D|w,\theta) (6)$$

4.1.2 Ranking Model. Collobert and Weston[7] use a pairwise ranking approach to capture the contexts of words for learning wordembeddings. It holds the similar idea with noise contrastive estimation but the optimizing objective is to assign a real word-context pair (w_i, h_i) a higher score than an artificial noise (w^n, h_i) by a margin. They minimize the following hinge loss function, where *T* is the training corpora.

$$loss_{cRank} = \sum_{(w_i,h_i) \in T} max(0, 1 - f_\theta(w_i,h_i) + f_\theta(w^n, h_i)) (7)$$

The scoring function $f_\theta(w, h)$ is achieved with a feed forward neural network. Its input is the concatenation of the current word $w_i$and context words h_i, and the output is a linear layer with only one node which stands for the compatibility between *w* and *h*. During training, an artificialnoise $w^n$is randomly selected over the vocabulary under auniform distribution.

4.2 Calculation sentiment polarity

In this section, we present the approach to encode sentiment polarity of sentences in sentiment embeddings. We describe two neural networks including a prediction model and a ranking model to take considerations of sentiment of sentences.

4.2.1 Prediction Model. An illustration of prediction model with binary sentiment categories (positive and negative) is shown in Figure 2(a). It contains five layers, namely *lookup → linear→ hTanh → linear→ softmax*. The input is a fixed-lengthword sequence $\{w_{i-c}, w_{i-c+1}, \ldots, w_i, \ldots w_{i+c-1}, w_{i+c}\}$,where $w_i$is the current word and *c* is window size. Lookup,linear and *hTanh*layers are described in Section 3.2.1. Theoutput of *hTanh*layer is used as features to predict thepositive and negative probabilities of input.

To predict the probabilities of positive and negative categories, we feed *hTanh*to a linear layer to convert the vector length to category number *C*which is 2 in the binary classification case.The parameters of the second linear layer are $Wl2 \in \mathrm{R}^{C\times len}$and $bl2 \in \mathrm{R}^{1\times C}$. We then add a *softmax*layer as the output layer to generate conditional probabilities over positive and negative categories. Letf^g (t) Є R ^{1*C} where

t= input t

C= number of sentiment polarity labels

For example, f^g (t) =[1,0] means sentence with positive polarity and f^g (t) =[0,1] means negative polarity. We use cross entropy between sentiment distribution and predicted distribution as the loss function of softmax layer. For the corpus T, the loss prediction for prediction model is defined as,

$$loss_{sPred} = -\sum_t^T \sum_{k=\{0,1\}} f_t^g\ (t).\, log(f_k^{pred}(t))(8)$$

4.2.2 Ranking Model. Now, we present an alternative of prediction model, which is a ranking model that outputs two real-valued sentiment scores for a word sequence with fixed window size. The basic idea of ranking model is that if the sentiment polarity of a word sequence is positive, the predicted positive score should be higher than the negative score. Similarly, if the sentiment polarity of a word sequence is negative, its positive score should be smaller than the negative score.

For example, if a word sequence is associated with twoscores[f_{pos}^{rank} ,f_{pos}^{rank}], then the values of [0.7, 0.1] can be interpreted as a positive case because the positive score 0.7 is greater than the negative score 0.1. Based on this consideration, we develop a neural network based ranking model, as illustrated in Figure 6 (b). As is shown, the ranking model is a feed-forward neural network consisting of four layers (*lookup → linear→ hTanh → linear→ softmax*.). Compared with the prediction model as shown in Figure 2 (a), the *softmax*layer is removed because it's objective does not require probabilistic interpretation.

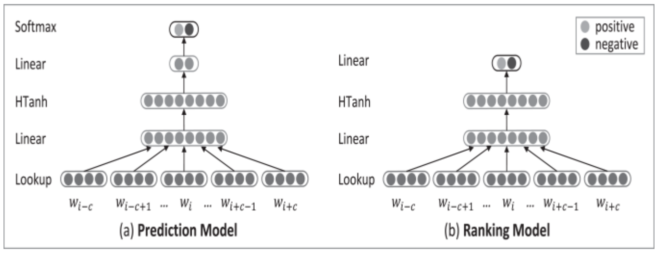


Figure 6. Layers in sentiment-aware word embedding

5. Experiment and Datasets

In this section, we explain about the data and compare the accuracy of difference word embedding methods on two datasets.

5.1 Data Collection

Table 1.Statics of the training datasets for sentiment-aware word embedding.

Dataset	#Positive	#Negative	#Total
IMDB	66,000	66,222	132,222
Twitter	637,728	665,432	1,303,160

We use two datasets for training the sentiment-aware specific word embedding separately. One is movie review data and the other is tweets.

5.2 Result and Analysis

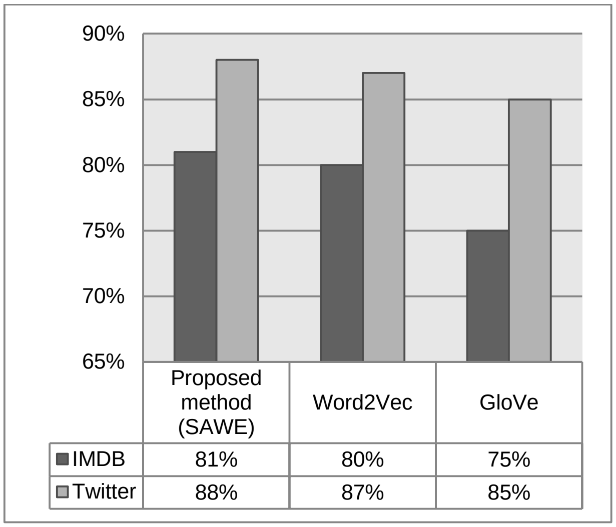


Figure 7.Accuracy of positive/negative classification with different word embeddings

Accuracies for different word embedding techniques are shown in Figure 7. Our proposed method gets higher accuracy than word2vec and GloVebecause it can capture for both context and sentiment information of text.Because of having the training data of Twitter dataset is 10 times of IMDB dataset, the different accuracies can be seen. The more data we can train, the higher accuracy we will get.

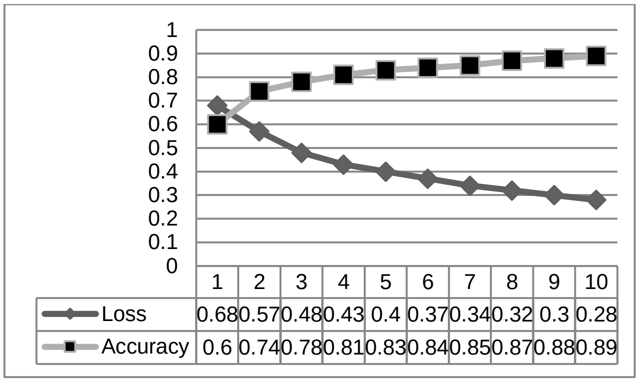


Figure 8.Loss and accuracy on IMDB dataset

Figure 8 explains about loss and accuracy on IMDB dataset. The loss function is one of the two parameters required to compile a model. Although there are many loss functions in neural networks, we use binary cross entropy in this system because it can classify on two classes.

6. Conclusion and Future Work

In this paper, we propose a sentiment-aware word embedding learning architecture for sentiment analysis. We encode the sentiment information into the continuous representation of words, so that it is able to separate good and bad to opposite of the spectrum. Finally, we prove that our method get higher accuracy than other word embedding techniques. Because of using sentiment-aware word embedding technique, it is effective for sentiment analysis than using normal word embedding techniques. For the future work, we aim to train on Burmese Text in different domain because language problem is also one of the challenges in sentiment analysis.

7. References

[1] Q.T.Ain, M.Ali, A. Riaz, A.Noureen, M.Kamran, B.Hayat and A.Rehman, “Sentiment Analysis Using Deep Learning Techniques: A Review”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.6, 2017, pp-424-433, 2017.

[2] P.Vateekul and T.Koomsubha, “A Study of Sentiment Analysis Using Deep Learning Techniques on Thai Twitter Data”, 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016.

[3] C.Li, B.Xu, G.Wu, S.He, G.Tian and H.Hao, “Recursive Deep Learning for Sentiment Analysis over Social Data”, IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI), 2014.

[4] L.Nio and K.Murakami, “Japanese Sentiment Classification Using Bidirectional Long Short-Term Memory Recurrent Neural Network”, The Association for Natural Language Processing, 2018.

[5] Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. “Efficient estimation of word representations in vector space.”In Proceedings of the 1st international conference on Learning Representations (ICLR 2013).

[6] PengFu,ZhengLin,FengchengYuan,WeipingWang,DanMeng, “Learning Sentiment Specific Word Embedding via Global Sentiment Representation”, Association for the Advancement of Artificial Intelligence, 2018, pp. 4808-4815.

[7] R. Collobert and J. Weston, “A unified architecture for natural language processing: deep neural networks with multitask learning,”in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 160–167

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” Journal of Machine Learning Research, vol. 3, pp. 1137–1155, 2003.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,”Journal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.

.

Simulation and Modeling in Software Approach

Functional Resonance Analysis Method on Road Accidents in Myanmar

Kyi Pyar Hlaing, Nyein Thwet Thwet Aung, Swe Zin Hlaing, Koichiro Ochimizu
University of Information Technology, Myanmar
kyipyarhlaing@uit.edu.mm, nyeinthwet@uit.edu.mm, swezin@uit.edu.mm,ochimizu@jaist.ac.jp

Abstract

In Myanmar, there are increasing numbers of road accidents in recent year. Because of road accident, 3480 people died in Myanmar in 2016 and approximately 10 people die from road accidents every day, according to Myanmar Organization Road Safety. Reasons for the accidents may include motorcyclists who are not wearing helmets, poor quality of road, driver behavior, mechanical failure of vehicles, violation of traffic rule, and bad weather. Functional Resonance Analysis Method (FRAM) is an accident analysis method providing a new concept for people to analyze accidents. FRAM can be applied by identifying functions with the detailed variability of functions, interpreting possible couplings of the variability and providing suggestions to manage the unexpected variability. This paper intends to propose basic FRAM model that analyzes the road accident. So it can provide a better understanding of accidents resulting from road accident. Based on this model, a case study was selected and the performance variability of function betweentwo functions by using FRAM model.

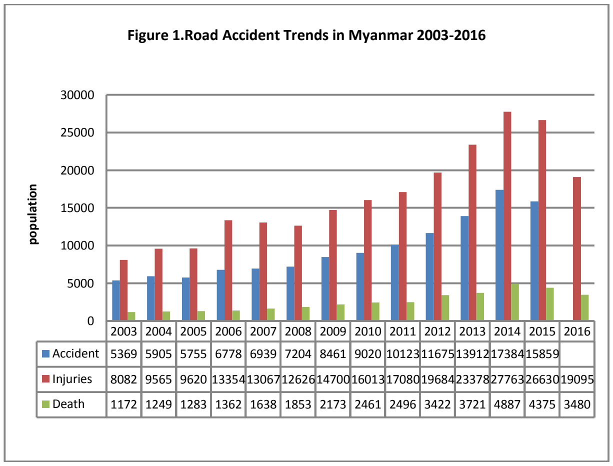
Keywords- Road Accident, Functional Resonance Analysis Method, Road Safety.

1. Introduction

Myanmar has the second highest death toll of road accidents in Southeast Asia, according to the Myanmar Organization for Road Safetythat quoted a WHO study. In Myanmar, road accidents are now major problems of country. Most of road accidents are caused by various defaults of drivers such as driving skill, drivers' judgment errors and violation of traffic rule, careless in driving and careless of pedestrian in walking.Some accidents are due to mechanical failure of the vehicles. All drivers are not aware of proper vehicles maintenance. According to an interview, one of the drivers discussed vehicles maintenance for highway express. He says that "to start a highway, they usually do general preparation and observation of vehicles' conditions only with their eye sight." Therefore they cannot know in detail about inside mechanical parts[6].

Functional Resonance analysis methods are promisingly used to derive potential accident scenarios. FRAM has many advantages in accident analysis. Therefore, FRAM is applied to investigate accident. FRAM focuses on the understanding of interactions andemergence phenomena in complex systems.FRAM can be applied by identifying functions with detailed information about how something is done, characterizing the variability of the functions, interpreting possible couplings of the variability, and providing suggestions to manage the unexpected variability.

Figure 1 shows the road accident trends in Myanmar, from 2003 to 2016.Fatalities increased to 1,853 in 2008, 2,496 in 2011, 3,721 in 2013,and then to 4,313 in 2014. The growth in the level of fatalities from 2013 to 2014 was 16%.The level of fatalities decreased in 2005-2016 [4].This paper aims to analyze road accidents between, driver and car.



	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
🟦 Accident	5369	5905	5755	6778	6939	7204	8461	9020	10123	11675	13912	17384	15859	
🟥 Injuries	8082	9565	9620	13354	13067	12626	14700	16013	17080	19684	23378	27763	26630	19095
🟩 Death	1172	1249	1283	1362	1638	1853	2173	2461	2496	3422	3721	4887	4375	3480

Figure1.Road Accident Trends in Myanmar, 2003-2016

2. Related work

Ana Gabriella Amorim [4]presents an application of functional resonance analysis method for an exploratory study of accidents at workplace where some types of improvisation took place. The FRAMexposed the recurrent weaknessor luck of control and supervision over the working process. As the result of using the FRAM method to explore the cases, it became clear that improvisations recognized as performance variation combined with other performance variations generating a functional resonance effect that was resulted in an accident.

In the FRAM-based analysis(FRAMA), the derivation of rules describing function variability (RFV) is highlighted to understand theinfluence of systemelements on each other, as well as to determine how the various performance of functionscan occur and aggregate. The RFV enables the analysis to be conducted by means of model checking(MC), and consequently facilitates exhaustive search based on the FRAM modeling, for potential performanceof the system functional model.The method FRAMA was applied to a typical ferry capsizal accident and the model checking resultsilluminate more details about the accident which causes than both the details provided in the officially-issuedinvestigation report and those produced by the current FRAM [3].

This paper proposed dynamic FRAM by combining thesystem analysis method of academic and industrial communityand tried to develop FRAM into a real-time analysis accordingto the time order. Dynamic FRAM is more effective to meetthe needs of the actual investigators, and it is also more effective in theanalysis of the relevant measures in the accident process. Theapplication of the method is illustrated by the case study of"7.23" Yong Wen serious railway accident [2].

This research investigates the linkage between highway road accidents and human rights issues in Myanmar.The research explores the reasons why there are so many accidents on the road and the underlying causes behind those accidents, the human rights issues emerged from those accidents and the required state obligations to promote right to life and right to health of passengers as parts of fundamental human rights [6].

FRAM provides an overview of how the system functions and how the emergent to explain the way that accidents happen. The research describes the cause of road accident by using FRAM to minimize accident in Myanmar.

3. Functional resonance analysis method (FRAM)

FRAM was originally developed for accident analysis (Hollnagel, 2004)[1,7]. However, it can also be used as an alternative approach to risk assessment and system modeling, bringing a new paradigm to manage and understand safety in complex socio-technical systems (Hollnagel, 2014)[1,7]. FRAM adopts a systemic and non-linear qualitative approach for system modeling and analysis by describing the normal performance variability within a socio-technical system. Variability in measured outcomes (such as performance, safety, etc.) is therefore introduced in socio-technical systems as an emergent property that is attributable to individual and collective/interacting human behavior during normal operations. Unexpected situations arise from higher degrees of this variability.

The FRAM method can be summarized with the following procedural steps:

Step 1: Identifying and describing the function

The premise of FRAM is the decomposition of the system into its functional entities, including the technical, operational, and organizational activities, which are involved in the day to day work of the system to succeed. Function can be characterized by the six different aspects or featuresbelow (Hollnagel, 2012), as shown in Fig. 2.

Input (I): that which the function processes or transforms orthat which starts the function.

Output (O): is the result of the function, either anentity or a state change.

Preconditions (P): conditions that must exist before a functioncan be carried out.

Resources (R):that which the function needs when it is carriedout (Execution Condition) or consumers' to produce the Output.

Time (T): temporal constraints affecting the function (withregard to starting time, finishing time or duration).

Control (C): how the function is monitored or controlled.

The six functional aspects are linked together to address the dependencies between the human technical activities during the specified scenarios as show in figure2 b.

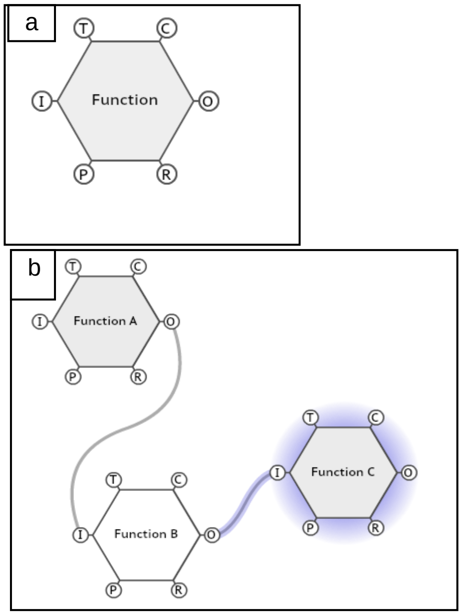


Figure 2 .(a)The six aspects characterizing a function (b)a demonstration of the functional dependencies are represented by the connecting lines.

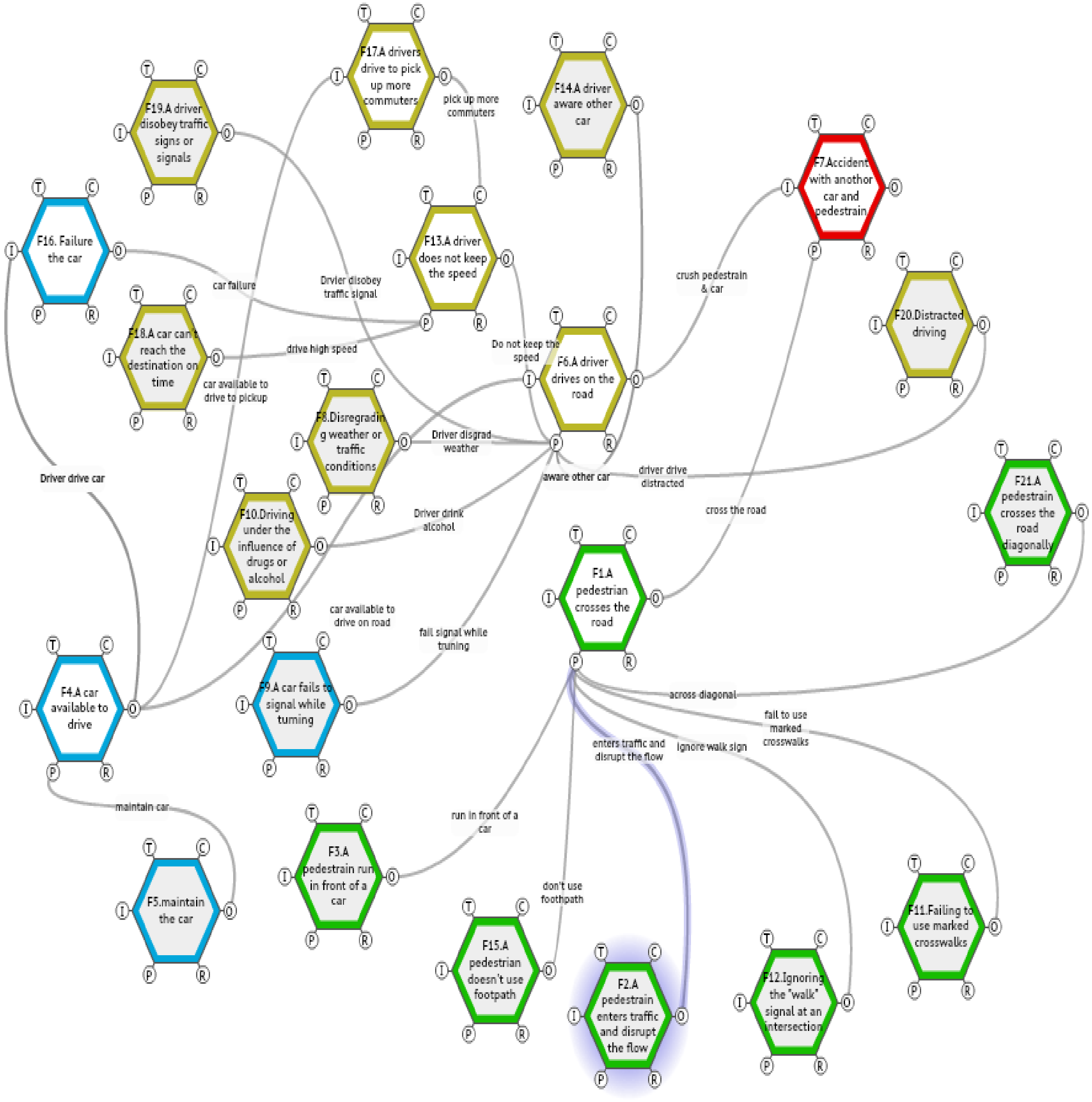

F17.A drivers drive to pick up more commuters
F14.A driver aware other car
F19.A driver disobey traffic signs or signals
F7.Accident with anothor car and pedestrain
F16. Failure the car
F13.A driver does not keep the speed
F20.Distracted driving
F18.A car can't reach the destination on time
F6.A driver drives on the road
F8.Disregrading weather or traffic conditions
F21.A pedestrain crosses the road diagonally
F10.Driving under the influence of drugs or alcohol
F1.A pedestrian crosses the road
F4.A car available to drive
F9.A car fails to signal while turning
F3.A pedestrain run in front of a car
F5.maintain the car
F15.A pedestrain doesn't use footpath
F2.A pedestrain enters traffic and disrupt the flow
F12.Ignoring the "walk" signal at an intersection
F11.Failing to use marked crosswalks
pick up more commuters
car failure
Drvier disobey traffic signal
crush pedestrain & car
drive high speed
Do not keep the speed
car available to drive to pickup
Driver disgrad weather
aware other car
driver drive distracted
Driver drive car
Driver drink alcohol
cross the road
car available to drive on road
fail signal while truning
across diagonal
fail to use marked crosswalks
enters traffic and disrupt the flow
Ignore walk sign
maintain car
run in front of a car
don't use foothpath

Figure4. FRAM model of Cause of road accident

Step 2: Characterizing the performance variability
The purpose of the second step is to characterize the variability of these functions. This should include both the potential variability of the functions in the model and the expected actual variability of the functions in an instantiation of the model.

Step 3: Aggregation of performance variability
The purpose of the third step is to look at specific instantiation of the model to understand how the variability of the functions may become coupled and determine whether this can lead to unexpected outcomes.

Step 4: Responding to performance variability

The purpose of the fourth and final step is to propose ways to manage the possible occurrence of uncontrolled performance variability that have been found by the preceding steps.

3.1. Variability

The FRAM comprises a number of functions that each describes an activity performed in theillustrated process.A key element of the FRAM is that it illustrates how different tasks areconnected or coupled to each other and how earlier activities can affect later activities by delayingor affecting the quality of the activity. The visualization of the FRAM illustrates that complexprocesses are difficult to describe in a linear way. Actions can vary or occur concurrently if thecircumstances or surroundings change [1]. Therefore, variability is investigated to provide anunderstanding of the couplings of functions.Variability can occur for different reasons [1]. First, functions can be affected by internalvariability caused by psychological or physiological factors. These can include stress, fatigue, well-being, decision-making ability, personal judgment and past experiences. Second, functions canvary due to the working environment in which they are carried out. This includes social factors suchas group pressure, social norms, relations with and expectations of co-workers, and the overallorganizational culture [1]. Finally, variability can evolve from upstream–downstream couplings [1]. Detecting such couplings is based on detecting potential variability and considering how itmay spread through the system and affect functions later in the process [1]. This view of processesand systems shows how variability emerges and how it is either amplified or dampened by actionslater in the process. Analysis of variability and couplings can help highlight the emergence ofunexpected outcomes; more importantly, it can describe how expected outcomes succeed despite variability in functions [1].

4. Functional Resonance Analysis Method on Road Accidents

In our propose system, there are four steps to analyze the cause of road accidents .In figure3; the first step is to identify the cause of road accident between car and driver, cars, and car and pedestrian. The second step identifies the actual or potential variability of each function according to the four basic groups (time/duration; force/distance/direction; object and sequence).Next step is to look at specific instantiations of the model to understand how the variability of the function may become coupled and to determine whether this can lead to unexpected outcome. The final step is to identify the key cause factors of road accident.

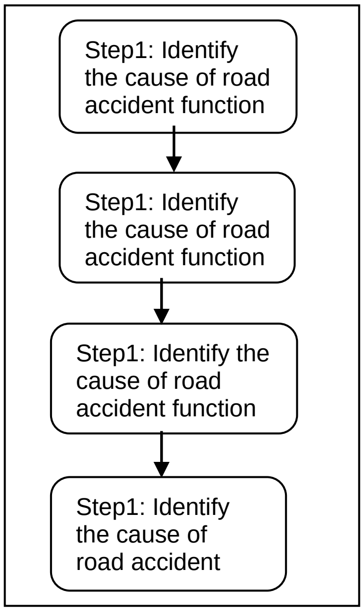

Step1: Identify the cause of road accident function
Step1: Identify the cause of road accident function
Step1: Identify the cause of road accident function
Step1: Identify the cause of road accident

Figure3.The overview of proposed system

Step1: Identify the cause of road accident function

The functions arecharacterized by the six aspects of Input, Output, Precondition, Resource, Time, and Control. The FRAM diagram aims to explain how normal performance variability may cause road accident. The diagram shows many functions when the driver and pedestrian cross the road.The six functional aspectswhich are linked together to address the FRAM diagram of Figure4 indicates the action of car, driver and pedestrian when they cross on the road.

Step 2: The identification of performance variability

Variability of a function is important for the process and can affect the later function in different way related to either time or quality, whether they varied internally,externally or due to couplings.This variability was caused by the default of drivers.Theyare absent to maintain andcheck the driver carefully. The possibility and consequences of internal variability were considered for functions in all sets in relation to how the variability affected the output of the functions. Upstream–downstream couplings will be described in further detail in the next step. In this case external variability was mainly considered in relation to the performance of technology function.

Another example of internal variability was observed at the function of adriver who does not keep the speed’. This variability can be causedbecause a car can’t reach the destination on time and a driver drives to pick up more commuters. In table 2, there are many reasons that show the performance variability of functions.

Table2. Identification of performance variability

Function	Function Name	Variability Description
F1	A pedestrian crosses the road	-A pedestriancrosses the road on time. -A pedestrian crosses the road later. -A pedestriancrosses the road earlier.
F4	A car available to drive	-A car is insufficient to drive -A car is ready to drive
F5	Maintain the car	The car has wrong action. The car has wrong object.
F6	A driver drives on the road	A driver drives wrong direction. A driver drives too fast. A driver drives too slow.
F10	Driving under the influence of drugs or alcohol	A driver drives wrong direction
F13	A driver does not keep the speed	-A driverspeed is too fast. -A driver speed is too slow.
F15	A pedestrian doesn't use footpath	Pedestrian can crush with car.

Step3. Aggregation of performance variability

In addition to variations caused by internal and external factors, function can vary because of upstream-downstream couplings. When attempting to understand a representation of reality, it is not necessary to know how variability may be combined. An example is presented in table3 and the function 'drives on the road'. In table 3, this function has many preconditions. The output of the precondition can be effect the downstream function.

Table3. The aggregation of performance variability

Function	Function	Variability Description
F4.A car is available to drive	F5.A driver doesn't check and maintain the car before driving	-Failure of the car -A car can't reach the destination on time. -Fail to signal while turning
F6.A driver drives on the road	F8.A driver disregards weather or traffic conditions F9.A car fails to signal while turning F10.A driver drives under the influence of drugs or alcohol. F14.A driver aware other car F19.A driver disobeys traffic signs or signals. F20.Distracted driving	-A driver can't drive the road safely -A driver can drive safely on the road
F1.A pedestrian crosses the road	F2.A pedestrian enters traffic and disrupts the flow F3.A pedestrian runs in front of a car F11.A pedestrian fails to use marked crosswalks. F12.A pedestrian ignores the "walk" signal at an intersection. F13.A driver doesn't keep the speed F15.A pedestrian doesn't use footpath. F21.A pedestrian crosses the road diagonally	A pedestrian can't across the road safely
F13. Driver does not keep	F16. If the car doesn't maintain and check, the car will fail. F17.If a driver drives to pick up more commuters. F18.A car can't reach the destination on time	-A driver doesn't keep the speed or keep the speed.
F16.Failure the car	F4.A car is not available to drive	-A driver doesn't keep the speed.
F17.A drivers drive to pick up more commuters	F16.A car isn't available to drive or doesn't available to drive.	-A driver drives the car with high speed to pick up more

		commuters.
F7.Accident with another car and pedestrian	F1.A pedestrian crosses the road F6.A driver drives on the road.	Accident can causebetween car and pedestrian

Step4. Identify the key cause factors

According to the functional resonance analysis method, the basic events in the process of accident are cleared. According to the cause study, table 4 shows some key causes factor of road factor and major damage of road accident. For example, failure of car's equipment is the key cause of road accident. Major damage that is failure of relevant components can't be replaced in time.

Table 4.Some of key causes of accident and related suggestions

Function	Major damage	Suggestion
-Failure of car's equipment	-Failure of relevant components that can't be replaced in time.	-Maintain the vehicle
- A driver didn't check a vehicle	- Can be accident during driving the car.	- Check the vehicle thoroughly
-A driver drinks alcohol	-A driver loses the ability to focus function and it's very dangerous when operating a vehicle.	-A driver shouldn't drink alcohol.
-A driver does not keep the speed	-A car can be crushed when a driver does not keep the speed.	-A driver should drive the normal speed
-A car can't reach the destination on time	- A car can be crushed when a driver drives with high speed	-A driver should drive the car to reach the destination on time.
-A driver running the red light	-They often cause side-impact collisions at high speeds	-To avoid a car accident, look both ways for oncoming car
-A pedestrians across the road diagonally	-It is very dangerous to the driver and the pedestrian	-To avoid a car accident, look both ways for oncoming cars

5. Evaluation

The aim of this research is to propose the basic FRAM model that analyzes on the road accident. So it can provide a better understanding of accidents resulting from road accident. Based on this model, a case study was selected that sow the the performance variability of function between driver and car by using FRAM model.

5.1 Case Study: YBS Accident

On July 7, 2017, Bus no. 55 and colliding with Bus no. 37that killed nine people and caused serious injuries to 30 other passengers. The accident was the worst since the Yangon Bus Service (YBS) started operating over one year in Yangon. Six people died on the spot and three people died at the hospital. The Bus no. 55 is 1998 model buses. The case study was used to extract the probability of road accident between cars, driver and car, car and pedestrian. According to the case study, figure 5 illustrates the model of the cause of road accident.

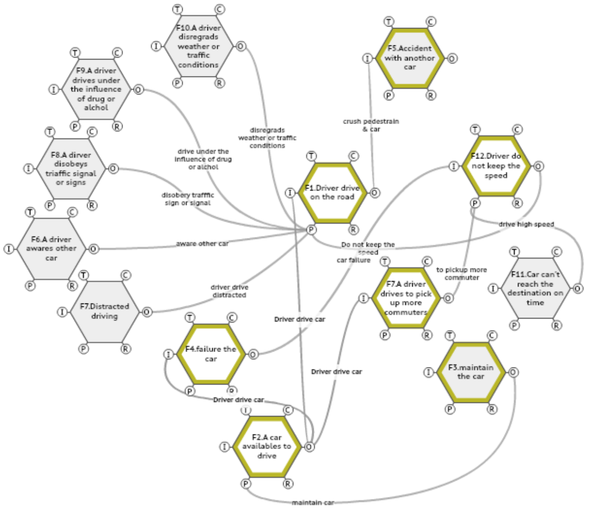

F10.A driver disregrads weather or traffic conditions
F5.Accident with anothor car
F9.A driver drives under the influence of drug or alchol
F8.A dirver disobeys triaffic signal or signs
F1.Driver drive on the road
F12.Driver do not keep the speed
F6.A driver awares other car
F7.Distracted driving
F7.A driver drives to pick up more commuters
F11.Car can't reach the destination on time
F4.failure the car
F3.maintain the car
F2.A car availables to drive
disregrads weather or traffic conditions
crush pedestrain & car
drive under the influence of drug or alchol
disobery trafffic sign or signal
aware other car
drive high speed
Do not keep the speed
car failure
to pickup more commuter
driver drive distracted
Driver drive car
Driver drive car
Driver drive car
maintain car
T
C
I
O
P
R

Figure5. FRAM model of Cause of YBS' accident

5.2 Discussion on analyzing of accident

The traditional report for the accident is the failure of vehicle. The proposed method shows the road accident which can be caused because of the driver who didn't check the vehicle, keep the speed, and drive to pickup more commuters, car can't reach the destination on time, failure of car's equipment. According to the investigation result, all the solutions were depended on the variations of the function performance. Such performance variations combined in an expected way and resulted in adverse situation.The comparison between accident report and FRAM application exemplifies that FRAM process does have strength over the traditional analysis process used in the accident investigation.By using FRAM model, we can express the cause of road accident in detail. The output of upstream function can affect the downstream function.

6. Conclusion

In this research, we proposed basic FRAM model that shows the causes of road accident. In Myanmar, there arean increasing number of road accidents in recent year. Road accidents are now the major problem of the country. By investigating the cause of road, we proposed a basic FRAM model which shows the interaction between cars, driver and car, and car and pedestrian. Based on the model, a case study was used to analyze the possible cause of road accident. By comparing the traditional accident report and proposed FRAM method there can be hidden causes of road accident. Method is applied to analyze the cause of road accident in Myanmar to minimize accidents. The proposed system has planned to consider more reliable in the future.

7. References

[1] Hollnagel E. FRAM: the functional resonance analysis method: modeling complex socio-technical systems. Ashgate Publishing, Ltd; 2012.

[2] Fang LIU, Jin TIAN,2017.**“**Dynamic analysis of FRAM”,The Second International Conference on Reliability Systems Engineering (TCRSE 2017).

[3]Jin Tian , Juyi Wu, Qibo Yang, Tingdi Zhao,2016 FRAMA:” A safety assessment approach based on Functional Resonance Analysis Method”,Safety Science 85 (2016) 41–52.

[4]Ana Gabriella Amorim*, Claudio M.N.A.Pereira,”Improvisation at workplace and accident causation - an exploratory study”, 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

[5] http://www.searo.who.int/myanmar/en/

[6] Human Rights Issues on Highway Road Accidents in Myanmar.

[7] Hollnagel, E., 2004. “Barriers and Accident Prevention”. Ashgate, Aldershot.

Performance Analysis of Layer Partition-based Matching Algorithm in Data Distribution Management

Nwe Nwe Myint Thein
University of Information Technology, Myanmar
nwenwemyintthein@uit.edu.mm

Abstract

High Level Architecture (HLA) is architecture for reuse and interoperation of simulations. In HLA paradigm, the Runtime Infrastructure (RTI) provides a set of services. Data Distribution Management (DDM) service reduces message traffic over the network. The evolution of the DDM service in HLA provides solutions to problems by using filtering mechanism that is suitable for large-scale simulation. These services rely on the computation of the intersection between "update" and "subscription" regions. When calculating the intersection between update regions and subscription regions, the higher computation overhead can occur. Currently, there are several main DDM filtering algorithms. The paper analyzes the performance of layer partition-based algorithms (LPM) for the matching process based on the different overlapping rate. The LPM algorithm provides the more definite matching area between update region and subscription region. The LPM algorithm guarantees low computational overheads for matching process of higher overlapping rate between the regions and reduce the irrelevant message among federates.

Keywords- HLA, DDM, Layer Partition-based, Matching Algorithm, Modeling and Simulation, Distributed System

1. Introduction

Efficient data distribution is an important issue in large-scale distributed simulations with several thousands of entities. The broadcasting mechanism employed in Distributed Interactive Simulation (DIS) standards generates unnecessary network traffic and is unsuitable for large scale and dynamic simulations.

DDM is a set of services defined in HLA to distribute information in distributed simulation environments. HLA's Run Time Infrastructure (RTI) is a software component that provides commonly required services to simulation systems. There are several groups of services, which are provided by RTI to coordinate the operations and the exchanges of data between federates (simulations) during a runtime execution. The interaction of object instances support by the function of RTI, which is similar to a distributed operating system. A simulation platform implements data distributed management in war game, airport modeling and simulation, air traffic control system and public transportation domain.

The remainder of the paper organizes as follows. Section 2 describes HLA issues relevant to data distribution. The previous algorithms for DDM matching methods explain in section 3. Section 4 represents the LPM algorithm for DDM. Section 5 presents the performance analysis of the system. Finally, section 6 offers conclusion.

2. Overview of DDM in HLA

DDM utilizes an N-dimensional coordinate system called a routing space to represent, for example, a geographical area. Federates express their interest by defining subscription regions that characterize the information they are interested in receiving. Each message is associated with a publication region to characterize the content of the message. If an overlap detect between a message publication region and a subscription region, the message will send to that subscribing federate. The main role of DDM is to reduce the volume of data exchanged through the matching process during a federation. Figure 1 shows the sample 2-deimensional routing space with four subscription regions and five publication (Update) regions.

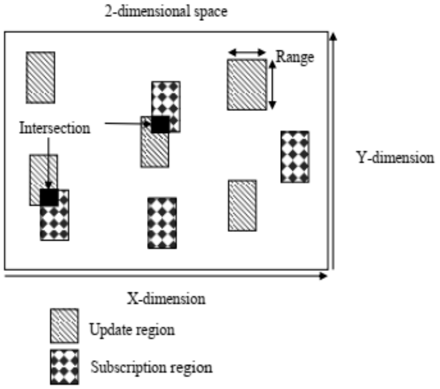


Figure 1: Example of region intersection in the 2-dimensional space

Table 1 presents the definitions of terms used in this paper, which originated from the HLA. [7]

Table 1: Terminology Definition in the DDM

Terminology	Definition
Dimension	A named coordinate axis with non-negative integers.
Multidimensional space	A coordinate system whose dimension is d (where d is a fixed natural number)
Range	A continuous semi-open interval on a dimension (lower bound, upper bound)
Region	A set of ranges for any given dimension.
Update region	A specified set of region instance for which is associated by a publishing federate.
Subscription region	A specified set of region instance for which is associated by a subscribing federate.
Overlap	All ranges of dimensions that are contained in the update region and subscription region put one upon another pairwise.
Intersection	An existence when the corresponding region sets overlap.
Matching process	A process to calculate the intersection between update and subscription regions.

3. Matching Algorithms in DDM

3.1. Region-based Algorithm

The region-based algorithm checks all the pairs of regions until an intersection found for each pair of update region and subscription regions or the end of the regions list reached. The implementation of this algorithm is straightforward, but the performance is varying greatly. [5] If there is N update regions and M subscription regions. “There are N*M pairs to check in the worst case. [7]”

3.2. Grid-based Algorithm

In the grid-based approach, the routing space partition into a grid of cells. Each region mapped onto these cells. If a subscription region and an update region intersect with the same grid cell, they assumed to overlap with each other. [9] Although the overlapping information is not exact, the grid-based algorithm can reduce the computation complexity than the region-based algorithm. [8] The amount of irrelevant data communicated in the grid-based filtering depends on the grid cell size, but it is hard to define the appropriate size of grid cells. [7]

3.3. Hybrid Approach

The hybrid approach is an improvement approach over the region-based and the grid-based approaches. The matching cost is lower than the region-based approach, and this advantage is more apparent if the update frequency is high. It also produces a lower number of irrelevant messages than that of the grid-based approach using large cell sizes. [9] The major problem is that it has the same drawbacks as the grid-based approach: the size of the grid cell is very crucial to the behavior of the algorithm. [6]

3.4. Sort-based Algorithm

The sort-based algorithm used a sorting algorithm to compute the intersection between update and subscription regions. [5] However, the sort-based algorithm's performance degraded when the regions are highly overlapped and it needed to optimize the sorting data structure for the efficient matching operation. [10]

3.5. Binary Partition-based Algorithm

The binary partition-based matching algorithm takes a divide-and-conquer approach similar to the one used for the quicksort.

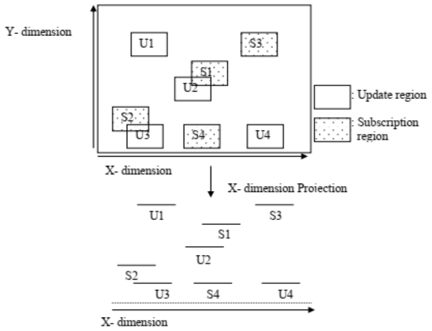


Figure 2: Dimension Projection with the X Dimension

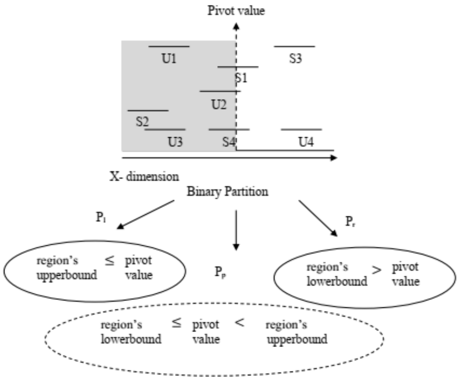

Pivot value
U1
S3
S1
U2
S2
U3
S4
U4
X- dimension
Binary Partition
P_l
P_r
P_p
region's upperbound ≤ pivot value
region's lowerbound > pivot value
region's lowerbound ≤ pivot value < region's upperbound

Figure 3: Binary Partition into Three Partitions, Pl, Pp and Pr

This approach consists of two main processes, the repetitive binary partitioning process, and the matching process. In the binary partitioning process, the algorithm recursively divides the regions into two partitions that entirely cover those regions. Second, in the matching process, the algorithm uses the concept of an ordered relation, which represents the relative location of partition. It easily calculates the intersection between regions on partition boundaries and does not require unnecessary comparisons within regions in different partitions, which are located in the ordered relation of partition. The process of algorithm describes in figure 2 and figure 3. The binary partition-based algorithm is not the best choice when the overlapping rate is relatively low. [7]

4. Layer Partition-based Matching Algorithm (LPM)

The Layer partition-based matching algorithm supports to search the overlapping information for data distribution management of HLA. It executes in dimension by dimension. This algorithm accepts all regions in the routing space. It also generates all regions randomly. Then it sends these regions to the Layer partition-based matching algorithm. The final overlapping information produces by observing the result of two matrixes for two-dimensional routing spaces. The LPM algorithm complete when all dimensions are covered. The detail instruction is state in figure 4. [3]

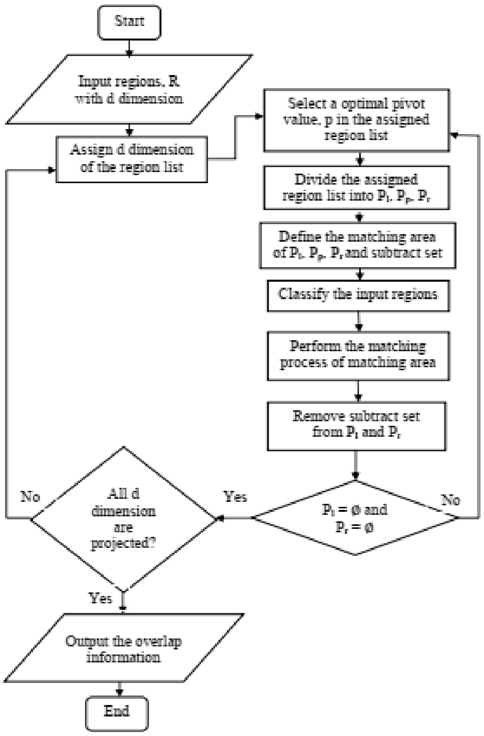

Start
Input regions, R with d dimension
Assign d dimension of the region list
Select a optimal pivot value, p in the assigned region list
Divide the assigned region list into P_l, P_p, P_r
Define the matching area of P_l, P_p, P_r and subtract set
Classify the input regions
Perform the matching process of matching area
Remove subtract set from P_l and P_r
$P_l = \emptyset$ and $P_r = \emptyset$
No
Yes
All d dimension are projected?
No
Yes
Output the overlap information
End

Figure 4: The LPM Dimension Algorithm

The LPM algorithm firstly chooses the optimal pivot to define the matching area. The efficiency and performance of the divide and conquer approaches depends on the choice of the pivot value. Some algorithms choose the middle point as the pivot value. The algorithm accepts the projected regions list and select one point of that list as the pivot value. At that point, the most subscriber regions and updater regions are converged in the projected regions list. In figure 5, the optimal pivot algorithm decides the optimal pivot value instead of the middle point. [2]

To define the exact matching area, a region distribution detection algorithm mainly used in the first layer of layer partition-based matching algorithm. The LPM algorithm firstly calculates the regions distribution. Then, the partitioning among regions performs based on the result of choosing pivot based on region detection and defines the matching area that entirely covers all regions, which need to match with regions at pivot point. The algorithm guarantees low computational overheads for matching process and reduces the irrelevant message among federates. [1]

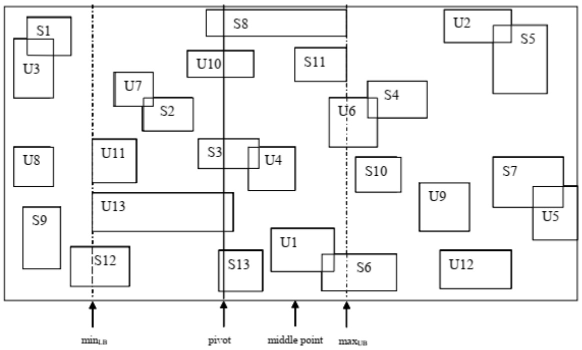


Figure 5: Sample Routing Space with 13 updaters and 13 subscribers

Table 2: Performance Analysis for Pivot Choosing

No. of updaters	No. of subscribers	% of overlap region	No. pivot choosing	
			Binary Partition-based Algo	Layer Partition-based Algo
13	13	0	32	15
13	13	7	32	15
13	13	15	32	15
13	13	23	32	15
13	13	30	32	15
13	13	38	32	15
13	13	46	32	15
13	13	53	32	15
13	13	61	32	15
13	13	69	32	15
13	13	76	32	15
13	13	84	31	16
13	13	92	31	16
13	13	100	30	16

Table 3: Performance Analysis for Matching

No. of updaters	No. of subscribers	% of overlap region	No. pivot choosing	
			Binary Partition-based Algo	Layer Partition-based Algo
13	13	0	44	20
13	13	7	44	20
13	13	15	44	21
13	13	23	43	20
13	13	30	43	22
13	13	38	43	20
13	13	46	43	20
13	13	53	44	22
13	13	61	43	22
13	13	69	42	20
13	13	76	42	22
13	13	84	42	21
13	13	92	42	21
13	13	100	43	21

The LPM algorithm promises the lower number of pivot point choosing. It also reduces the number of matching process between the updater regions and the subscriber regions of the routing space. The analysis of the LPM with 13 updaters and 13 subscribers describe in Table 2 and Table 3. The area of the routing space is 100*60. We assume that the number of pivots choosing for worst case is 50 for X dimension and 30 for Y dimension. We also define the number of matching between the two kinds of regions is 2*(13*13). The LPM algorithm reduces the half of matching process by defining the exact matching area. It also assures the lower number of pivots choosing for partition the routing space. [2]

In the second layer, the specific decision of the region's selection performs to calculate the matching data between the three sets. This layer also supports the subtracted region lists. These lists subtract from the input regions set for next matching calculation. The classification of regions carries out in the region classifier algorithm. The actual matching between the updater regions and the subscriber region execute in intersection calculation algorithm. The subtracted region lists use to reduce the next calculation. [3]

The Layer partition-based matching algorithm consider in 2-dimensional routing space as three different ways. The first method uses the same number of input regions in each dimension. The final overlapping information can get by using AND operation between the overlapping result of two dimensions. [4]

The number of input regions in each dimension of second method is different. The input regions in Y dimension depends on the overlapping result of the X dimension. If some of input regions in X dimension are not overlapped, they cannot include in the input regions of Y dimension. The final overlapping information can produce the result of overlapping matrix in Y dimension. [4]

The third method is piggyback the result of matching result of the X dimension. Before the making decision for Y dimension, the LPM algorithm needs to check the matching result of X dimension. The final overlapping information decides by Y dimension without combing the overlapping results of two dimensions. Three methods of the LPM algorithm will complete when all dimensions are covered. [4]

5. Performance Analysis

For the matching algorithms of DDM, the impact of network speed on the algorithm does not care and actually, there are no messages transferred in the network in all of the approaches. Thus, a single computer used to make experiments. As the performance of the DDM execution time for the matching process is measured with Microsoft Windows 8 with 2.90GHz Intel(R) Core (TM) i7 CPU and 8GB memory. One of the important experimental parameters is the number of regions. The overlap rate defines as the proportion of the scene volume occupied by the regions. Therefore, we define the overlap rate as shown in equation 1:

$$\text{overlap rate} = \frac{\sum \text{area of regions}}{\text{area of space}}$$

where $\sum$ area of regions = number of regions * high of region * width of region. If the routing space is 100*100 and one region is 1 * 1, where the number of regions is fixed at 100, the overlap rate is

$$0.01 \;=\; \frac{100 * (1*1)}{100 * 100}$$

5.1 Theoretical Analysis on Computational Complexity

To analyze the computational complexity of the LPM algorithm, we suppose that there are N regions with the number of dimensions, d=2 in the multidimensional space. The optimal pivot algorithm requires O(N) computation for the size of region is M. We assume that the size of region and the number of dimensions is constant. The first layer partition algorithm requires O(N) computation. The total number of recursions for matching algorithm requires O (log N) computation.

Moreover, the second layer partition algorithm also needs O(n) computation and the matching process of comparing the intersection of regions between partitions requires O (n^2) computation (where n is the number of regions in each partition). The complexity of the intersection calculation procedure is proportional to n. It seems that the most important points are the exact matching partitions. It is obvious that the number of regions, n, is a determinant factor. Because the overlap information of all regions obtain by the pivot partition, it is not necessary to compare their overlap information in the left and right partitions of pivot partition. Therefore, the computational complexity of the LPM algorithm is n^2 x N x O (log N) computation. If the number of regions, n, is normally very small in a large-scale spatial environment, so the LPM algorithm should be very efficient. Therefore, the actual computational complexity depends on how the exact matching partition well achieved.

5.2 Performance of DDM Algorithms

The regions distribute randomly across the routing space 10000 * 10000. The number of regions is differing from 1000 regions to 15000 regions.

5.2.1 Performance Analysis of LPM Using Same Size Region. The performance analysis based on the overlap rate 0.01, 0.1 and 1. The figure 6 and 7 shows the execution time for the matching process in four other algorithms and LPM when the overlap rate is 0.01. The figure 8 and 9 shows the execution time for 0.1. For the overlap rate 1 is shown in figure 10 and 11.

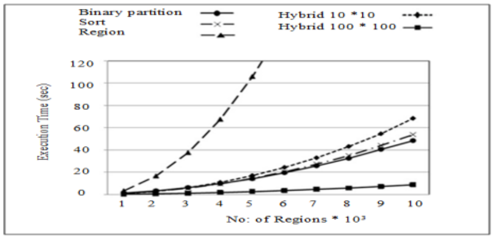


Figure 6: Performance of Overlapping Rate 0.01

It seems that the hybrid approach with 100 * 100 grid cells always has the best performance. The binary partition-based matching algorithm outperforms the other matching algorithms when the overlap rate is 0.01.

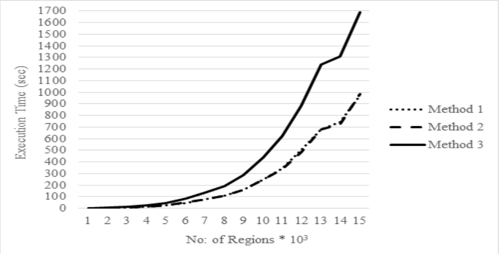


Figure 7: Performance of Overlapping Rate 0.01 of LPM

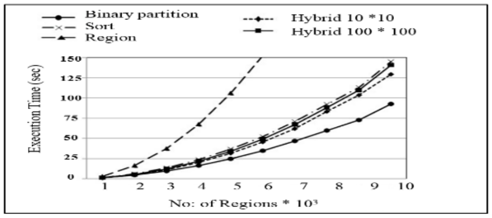


Figure 8: Performance of Overlapping Rate 0.1

In figure 8, the computational overhead of hybrid approach degrades significantly when the number of regions is higher.

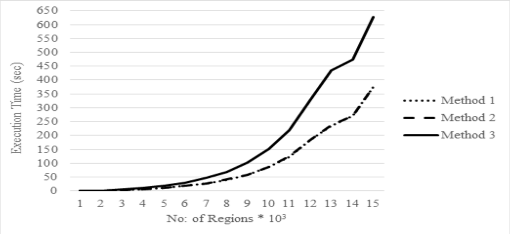


Figure 9: Performance of Overlapping Rate 0.1 of LPM

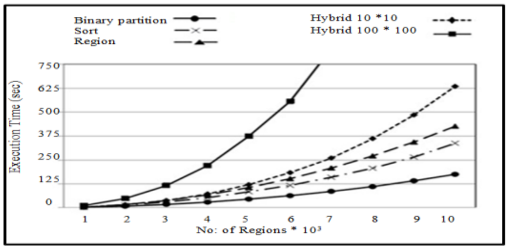


Figure 10: Performance of Overlapping Rate 1

With the overlap rate is 1, the binary partition-based algorithm performs well. On the other hand, the sort-based algorithm performs better, except the binary partition-based algorithm in the overlap rate is 1. When the number of regions increases and the overlap rate is high, the performance of the region-based algorithm becomes increasingly better than the other overlap rate. From all of the figures, we know that the hybrid approach with 100 * 100 grid cells has an extremely big computational overhead for the matching process.

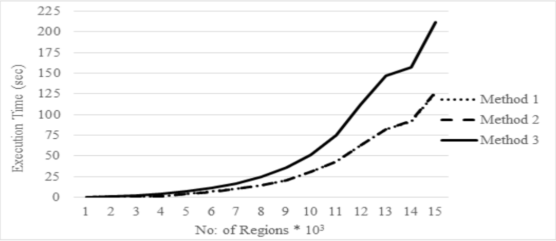


Figure 11: Performance of Overlapping Rate 1 of LPM

For the Figure 7, 9 and 11, the performance of the LPM algorithm analyze on same size regions, which generate randomly. The three methods are of LPM is not the best choice when the overlapping degree is relatively low, but it has the advantage of the matching time when the overlap rate is high. The performance of the first method and the second method are nearly the same. The input region list for second dimension cannot affect the overall matching process. The best method of LPM is the first method. According to the analysis results, it is proved that the execution time of same size regions can be reduced about two third than the previous matching algorithms for the overlapping degree 1.

5.2.2 Performance Analysis of LPM Using Different Size Regions. To define the size of each region base on overlap rate 0.01, 0.1 and 1. To generate the different size region upon the routing space, the equation 2 is used.

$$\text{region size} = \sqrt{\frac{\text{area of space * overlap rate}}{\text{number of regions}}} \qquad (2)$$

The number of regions is the fifteen different sizes from 1000 to 15000. All methods of LPM algorithm are more efficient than the existing matching algorithms of DDM at any overlapping degree using different size regions. The main advantage of this algorithm is its support for scalability very well, when the overlapping degree is large.

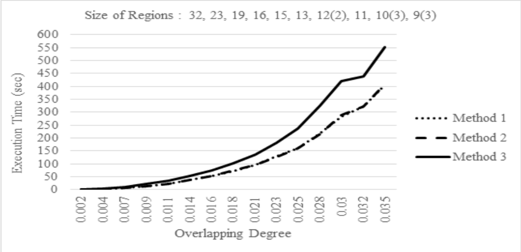


Figure 12: Different Size of Regions Generated by Overlapping Rate 0.01

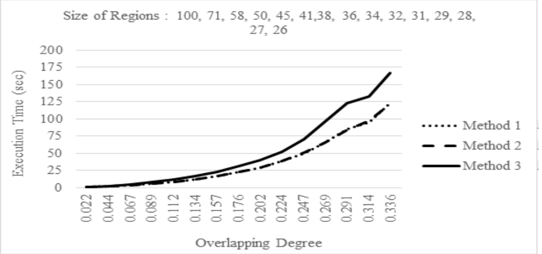


Figure 13: Different Size of Regions Generated by Overlapping Rate 0.1

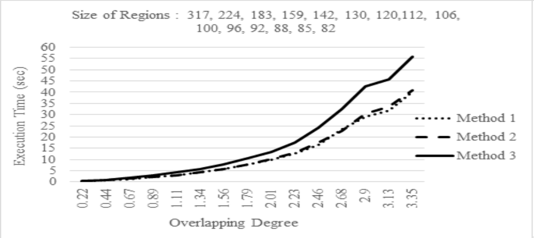


Figure 14: Different Size of Regions Generated by Overlapping Rate 1

6. Conclusion

The layer partition-based matching algorithm is very useful and efficient for the different size of the regions. The LPM algorithm can improve the efficiency and performance by the right choice of the optimal pivot algorithm. The matching process can decrease the number of comparing regions between the pivot partition and left partition and pivot partition and right partition. The number of regions in the projected region list can reduce over and over again by using subtract list. The LPM algorithm does not need the partitioning to cover all regions. According to the analysis result, the LPM algorithm of method one is better than the previous algorithms when the overlapping degree is higher. It supports the best matching result with different region sizes. The final overlapping result can obtain in a timely and efficient manner.

7. References

[1] M. T. Nwe Nwe and T. Nay Min, “Optimization of Region Distribution Using Binary Partition-based Matching Algorithm for Data Distribution Management”, International Journal of Engineering Research and Technology (IJERT), Vol. 2 Issue 2, February 2013, pp. 1582-1587.

[2] M. T. Nwe Nwe and T. Nay Min, “Dynamic Pivot for Layer Partition-based Matching Algorithm of DDM based on Regions Distribution”, Proceedings of the 4th International Conference on Science and Engineering (ICSE), Yangon, Myanmar, December 9-10, 2013.

[3] M. T. Nwe Nwe and T. Nay Min, “Layer Partition-based Matching Algorithm of DDM”, Proceedings of the 3rd International Conference on Computational Techniques and Artificial Intelligence (ICCTAI), Singapore, February 11-12, 2014.

[4] M. T. Nwe Nwe and T. Nay Min, “Layer Partition-based Matching Algorithm of DDM based on Dimension”, International Journal of Advanced Computer Technology (COMPUSOFT), Vol. 3 Issue 4, April 2014, pp. 691- 695.

[5] R. Come, T. Gary and S. C. Tay, “A Sort-Based DDM Matching Algorithm for HLA”, ACM Transactions on Modeling and Computer Simulation, Vol. 15, Issue 1, January 2005, pp. 14-38.

[6] T. Gary, Z. Yusong and A. Rassul, “A Hybrid Approach to Data Distribution Management”, Proceedings of the 4th IEEE International Workshop on Distributed Simulation and Real-Time Applications, San Francisco, CA, August 17-25, 2000, pp. 55–61.

[7] A. Junghyun, S. Changho and G. K. Tag, “A Binary Partition-based Matching Algorithm for Data Distribution Management”, Proceedings of Winter Simulation Conference (WSC), Phoenix, AZ, December 11-14, 2011, pp. 2723-2734.

[8] A. Rassul, M. Farshad and T. Gary, “Optimizing Cell-size in Grid-based DDM”, Proceedings of the 14th Workshop on Parallel and Distributed Simulation, Bologna, Italy, May 2000, pp. 93–100.

[9] G. Tan, R. Ayani, Y. Zhang and F. Moradi, “Grid-based data management in distributed simulation”, Proceedings of the 33rd Annual Simulation Symposium, Washington, DC, April 2000, pp.7–13.

[10] J. Yu, R. Come and T. Gary, “Evaluation of a Sort-based Matching Algorithm for DDM”, Proceedings of the 16th Workshop on Parallel and Distributed Simulation, Washington, DC, May 2002, pp. 68–75.

Myanmar Semantic Information Retrieval Using Self Organizing Map with Global Vector – MyanSeM

Thiri Haymar Kyaw, Thinn Thinn Wai, Thinn Mya Mya Swe
University of Information Technology
thirihaymarkyaw@gmail.com, thinnthinnwai@uit.edu.mm, thinnmyamyaswe@uti.edu.mm

Abstract

Nowadays, explosive growing the resources with Myanmar language on the Internet, the information retrieval (IR) for Myanmar web pages has increasingly important.The proposed system presents an effective semantic retrieval approach based on parallel Self Organizing Map (SOM), which uses document associationinstead of Euclidian distance in distance calculation, and GloVe (Global Vector for Word Representation) for word co-occurrence. *The Self Organizing Map (SOM) has been a promising method for document clustering and word sense disambiguation. This approach uses the parallel training the separate parts of SOM for document clustering, then combine, and re-cluster the documents.During the training of the parts of SOM, global vector is used for appearance of word co-occurrence and then combines the word categories based on semantic sense.Although many researchers have researched the various semantic information retrieval approaches, they have not yet adapted to retrieve the semantic information of Myanmar words and sentences. This approach can retrieve most semantically relevant web documents. This does not take too long time for SOM training because of parallel GPU approach.*

Keywords-Semantic Web, Self-Organizing Map, Information Retrieval.

1. Introduction

The World Wide Web serves the vastly distributed information services for every kind of information such as news, advertisements, blogs, customer relationship management, online learning, e-government, e-commerce, health services, context awareness services, etc. with custom languages. Myanmar language is same as the other languages like English. Some words are polysemy or homonyms and some words are synonyms. In addition, Myanmar words are un-segmented words and the delimitation of words is based on the typing of the user. Sometimes, the user does not type the word segments properly. The well-known search engines segment the Myanmar words according to the space delimitation and do not consider semantically related Myanmar words. They cannot retrieve the user-satisfied result. They retrieve a large number of irrelevant documents that are unable to meet the user's request. Analyzing the semantic meaning of words in user’s query performs semantic information retrieval. The proposed system focuses on the Myanmar web pages for retrieving most relevant results.

There have been several researches applying Self Organizing Maps (SOMs) for word sense disambiguation [10], web log clustering [4, 5], document clustering and information retrieval [1, 3, 6, 7, and 10]. The Self-Organizing Feature Map (SOFM or SOM) is a clustering and data visualization technique based on a neural network proposed by Teuvo Kohonen [18, 19]. SOM is a model of unsupervised machine learning and an adaptive knowledge representation scheme. SOM consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. The self-organizing map describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector [3].

The previous SOM based clustering approaches can organize the document map from word category [21] or WordNet [17] and enhance the information retrieval [10, 21]. Most of them are not consider for the semantic relation of words or words co-occurrence. They use one SOM for training the whole document collection and it takes huge amount of time and memory. Separating the parts of SOM and running in parallel [15] is the more promising approach for training the document collection. For those purpose of word representation, word vector approaches such as SVD based methods, continuous bag of words (CBOW), Skip-Gram, and Global Vectors (GloVe) [9]. The previous approaches are not convenient for the semantic meaning of the Myanmar words in the query and the documents very well. The proposed MyanSem method provides a deep SOM with GloVe for the retrieval of Myanmar documents.

2. Related Work

Pushpa et al. compared web page recommendation systems using K-means and Self Organizing Map [3].

Both the methods used historical browsers data for search key words and provided users with most relevant web pages. All users' click-through activity such as number of times he visited, duration he spent, and several other variables were stored in database. These systems used this database and process to cluster and rank them. The obtained results showed that the Self Organizing Map technique produced the most relevant results for a particular query word compared to K-means technique.

Tarek F. Gharib, et al. proposed the semantic text document clustering approach that using WordNet lexical and Self Organizing Maps [17]. This approach used the WordNet to identify the importance concepts in the document. The SOM is used to enhance the effectiveness of document clustering algorithms. This approach took the advantages of the semantics available in knowledge base and the relationship between the words in the input documents. They have two reasons for using SOM that it is topologically preserving and clustering is performed non- linearly on the given input data sets. The topologically preserving property allows the SOM applied to document clustering, to group similar documents together in a cluster and organize similar clusters close together unlike most other clustering methods.

Xia Lin [21] proposed a Self-Organizing semantic map for retrieval of AI literature. The purpose of this system was to conceptualize art information retrieval approach, which used traditional search techniques as information filters and the semantic map as a browsing aid to support ordering, linking, and browsing information gathered by the filters.

Peter Gajdos and Pavel Moravec presented a simple modification of classic Kohonen SOM; they called Global-Merged SOM. Their approach allowed parallel processing of input data vectors or partitioning the problem for all vectors from the training in reducing the memory consuming. The set of input vectors is divided into a given number of parts. The classic SOM is applied on every part. In the final phase, pre-selected potential centroids of data clusters are used as weight vector. Their algorithm can utilized the power of batch processing in all inner parts (PSOM) [15].

3. The Proposed System

The focus of the MyanSeM method is to cluster Myanmar document collections using parallel SOM, in order to consider dimension reduction by finding word co-occurrence based on global vector [9]. The proposed method also addressed for the polysemy words of the query sentence choosing the correct sense.

3.1. Preprocessing

First, the information retrieval systems make the preprocessing for the crawled web pages such as removing HTML tagsand identifyingthe main content blocks. After removing HTML tags, the preprocessing tasks such as stopword removal, stemming, and handling of digits, hyphens, punctuations are also required [2]. Myanmar language is un-segmented language like Chinese, Japanese, and Thai. The tokenization or segmentation process is more difficult than space-delimited languages like English. The part of speech tagging of Myanmar words based on the corpus or lexicon is needed for constructing the term-document vector.The system also segments the query sentence to terms or words. A document in the vector space model [2] is represented as a weight vector, in which each component weight is computed based on some variation of TF (Term Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency) scheme. In this paper, we do not explain detail about pre-processing and word segmentation.

3.2. Self-Organizing Map

Self-organization map (SOM) is an unsupervised learning method where the neural network organizes itself to form useful information [12]. Kohonen innovated this principle of topographic map formation [18, 19]. The SOM uses a set of neurons, regularly arranged in one dimension or two dimension rectangular or hexagonal grid, to form a discrete topological mapping of an input space. The architecture of two-dimensional hexagonal grid Kohonen self-organizing map is illustrated in Figure 1.

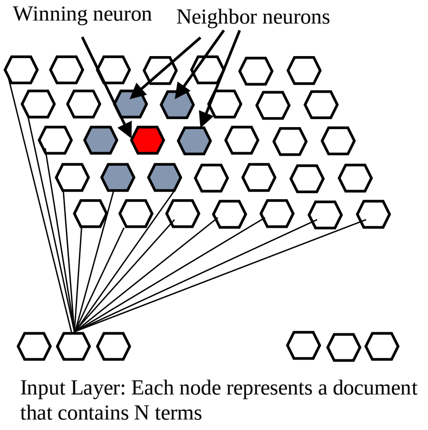

Winning neuron
Neighbor neurons
Input Layer: Each node represents a document that contains N terms

Figure. 1 Architecture of Self-organizing map

Continuous input values are presented sequentially in time through the input layer, without specifying the desired output. Each pattern is represented in the form of a vector, $x = (x_1, x_2, ..., x_n)$. Computational procedures of SOM [20] are summarized as follows:

Step 0. Initialize

- Small real random values to weights, $w_{ij}$for $i = 1$ to n and $j = 1$ to m.

- A neighborhood parameter, *h* and a learning rate, α, where $0 \leq \alpha \leq 1$

Step 1. Enter a new input vector, $x = (x_1, x_2, ..., x_n)$, to the input layer.

Step 2. Select winner neuron with the smallest distance to *x*.

Step 3. Modify Weights.

- Adjust the winner and its neighbors' weights according to the following formula:

$$w_j(t+1) = w_j(t) + \alpha(x(t) - w_{j(}t))$$

Step 4. Update learning rate α. Slowly reduce radius *r* at specified iterations.

Step 5. Continue from Step 1 to 4 until weights have stabilized.

The SOM algorithm always converges to a solution, i.e., that each of the winner weight vectors of the map converges to the mean of the data vectors for which it has been a winner, in a finite number of steps [18,19].

3.3. GloVe for Word Representation

In the *GloVe* model [9], the global corpus statistics are captured directly by the model. Let the matrix of word-word co-occurrence counts be denoted by *X*, whose entries X_{ij} tabulate the number of times word *j* occurs in the context of word *i*. Let $X_i = \sum_k X_{ik}$ be the number of times any word appears in the context of word *i*. Finally, let $P_{ij} = P(j|i) = X_{ij} / X_i$ be the probability that word *j* appear in the context of word *i* [9]. The simple example of Myanmar word co-occurrence probabilities is shown in Table 1.

Table. 1 Co-Occurrence Probabilities for Target Words တူ (hammer) and သတ် (kill) with Selected Context Words ထု (beat), ရိုက်(strike), စား(eat) and တော်စပ် (relate)

Probability and Ratio	k = ထု	k = ရိုက်	k = သေ	k = တော်စပ်
P(k| တူ)	3.4×10^{-3}	1.5×10^{-4}	0.3×10^{-4}	1.7×10^{-4}
P(k| သတ်)	2.9×10^{-3}	1.7×10^{-4}	4.1×10^{-4}	2.1×10^{-5}
P(k| တူ)/ P(k|သတ်)	1.17	0.88	7.3×10^{-2}	8.1

Consider two words *i* and *j* that exhibit a particular aspect of interest. The relationship of these words (တူ[hammer or chopstick],သတ်[kill],ထု[beat], ရိုက်[strike],စား[eat],တော်စပ် [relate]) can be examined by studying the ratio of their co-occurrence probabilities with various probe words, *k*. After doing vector difference, the highest occurrence probability ratio of words *i*, *j* to other work *k* is point out. For words k related to တူ but not သတ်, say k = တော်စပ်, we expect the ratio P_{ik}/ P_{jk} will be large. Similarly, for words *k* related to သတ် but not တူ, say *k* = သေ, the ratio should be small. For words *k* like ထု or ရိုက်,either these are related to both တူ and သတ်, or to neither, the ratio should be closeto one.

Theratio P_{ik} / P_{jk} depends on three words *i*, *j*, and *k*,the most general model takes the form,$F(w_i , w_j , \bar{w}_k) = P_{ik} / P_{jk}$ where $w \in \mathbb{R}^d$ are word vectors and $\bar{w} \in \mathbb{R}^d$ are separate context word vectors.

We trained our model on one lexicon based on 100 Myanmar news, commercial and blog web sites with 1 million tokens. We tokenizeand build a vocabulary of the 1000 most frequent words and then construct a matrix of co-occurrence counts. Currently, we omit the name entity recognition.

3.4. MyanSeM Algorithm

The proposed algorithm has two phases: Training phase and Query phase. In the training phase, it provides the clustering of Myanmar documents into related groups by caring the semantic categories and co-occurrence of words. In the query phase, it considers the semantic sense of the query sentence, maps to the most related clusters and retrieves the most relevant web pages.

Let *T* is the term vector and $T \in t_j$, where $j = 1$ to *N*. *N* is the number of terms in the document collection (*D*) and $D \in d_i$where $d_i$represents each document and $i = 1$ to *M*. *M* is the number of documents. Each term (Myanmar word) comes from the Myanmar Lexicon that includes the Myanmar word, Myanmar meaning (definition), and English meaning and example Myanmar sentences of all homonyms. Term-document matrix is filled with weights of terms in each document.

SOM is composed of neurons that are also called nodes. Each node represents one document (d_i). Weight vector of each document $d_i = [w_1, w_2, ..., w_n]$. Each weight $w_j$of term *j* in document *i* represents *TF*(term frequency) * *IDF* (inverse document frequency) [2].

The proposed approach uses the two-dimensional hexagon shape SOM. In this approach, we use parallel SOM based on GPU-based SOM [15] that divides the document vector into parts of the vector. The collection of document *D* is divided into *P* number of parts. One part contains *K* number of documents.The partition of document collection is based on the number of documents in the collection.In the sample test in Section 4, we divide four partition (*P*) on 100 documents and 25 documents in each partition.

The stepbystep procedure for MyanSem algorithm is as follows:

[Training Phase]

Step 1: SOM with GloVe

- Train SOM for clustering documents within P partition
- Choose one document d among K documents randomly as an input vector.
- Find the most associate document $d_i \in K$, $i = 1$ *to* K.

Association of two documents: $\frac{|d \cap d_i|}{|d|}$

where

$| d \cap d_i |$ means number of words (terms) contain in both document d and d_i,

$| d |$ means the number of words (terms) contain in document d.

- Extract terms that contains in both input document and the most associate document are called shared terms S.
- Construct the Global Vector for those shared terms and the other terms that contains separately in two documents.
- Determine co-occurrence words (terms) that have highest ratios. Union those term vectors.
- Train SOM as traditional SOM until convergence for K documents. Terms have already reduced by combining co-occurrence terms.
- Compose L number of cluster that contains most associate documents. Select the cluster centroids (winner) for all L clusters.

Run Step 1 in parallel for P partitions.

Step 2: Combine the clusters

- Take one cluster randomly and the centroid node (document vector) within this cluster acts as the input vector.
- Train the SOM for all clusters until no changes occur.

[Query Phase]

Step 3: Sense the Query

- Construct the query word vector
- Check the homonyms of words using the Myanmar wordnet like lexicon.
- Construct local co-occurrence vector of query words and map to the examples of word in lexicon.

Table. 2 Query Word Co-Occurrence Matrix Applies to the Example Sentences in Lexicon

	တူ (hammer)	ဖြင့် (with)	ထု/ရိုက် (beat/strike)
တူ (hammer)	0	e1, e2	0
ဖြင့် (with)	e1, e2	0	e2
ထု/ရိုက် (beat/strike)	0	e2	0

Let consider the simple example query “တူဖြင့်ရိုက်”. The word ″တူ″ has four different meanings (chopstick, hammer, relate, same) and the word ″ရိုက်″ has a synonym ″ထု″ that is already combine in the term vector. The examples are e1:“ခေါက်ဆွဲကိုတူဖြင့်စားသည်”, e2:“သံကိုတူဖြင့်ထုသည်” and e3:“တူတော်စပ်သည်” e4: “ဆင်တူသည်”.

In the query, word co-occurrence matrix shown in table 2, the example sentence of e2 is best match with the query. So, the correct sense of word (hammer) is selected.

Step 4: Map the Query to the document clusters

- The query vector enters as the input vector. The words that are not contains in the query sentence are filled with zero.
- Calculate the document similarity or association between the query vector and the cluster centroids.
- Choose the best match cluster and rank the documents within this cluster using query co-occurrence matrix applying these documents.

4. Discussion

We collect the 100 documents for training and the example sentences are shown in Figure 2. There are 50 total vocabulary (terms) used in the documents. The 50 × 100 document-term vector is separated into four 50 × 25 vectors. First we trains the SOM on first 50 × 25 vector with some epochs and construct the global vector for words (terms) within each cluster. Then we can reduce the 25 terms to 20 terms that have high co-occurrence probability. Then, we continue to train the SOM for 50 × 20 vector. As the above mentioned, we apply four SOM parallel and then combine after running the SOM until it reaches convergent. Finally, the semantically related documents are placed together within the same cluster. The sample clusters are shown in Figure 3.

When the user enters the query “တူဖြင့်ရိုက်”, first we determine the correct sense of every word in the query sentence as mentioned in Step 3. The query enters as an input to the combined SOM and chooses the winner and neighbors as the query results. We use the cosine of the angle (cosine similarity)[2] as the distance measure for choosing the winner of SOM as follows:

$$cosine(\mathbf{d}_j, \mathbf{q}) = \frac{\langle \mathbf{d}_j \bullet \mathbf{q} \rangle}{\| \mathbf{d}_j \| \times \| \mathbf{q} \|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$

Assume that *q* is the query vector *q* and $d_j$is the document vector, which is the centroid of each cluster. Most similar cluster centroid (winner node of SOM cluster) is selected. Finally, we construct the word vector of the query words (terms) and words (terms) in the documents within the winner cluster. The result of the above query is shown in Figure 4.

d1	ကလေးသည် ခေါက်ဆွဲကို တူနှင့် မစားတတ်ပဲ ဖြစ်နေသည်
2	ဂျပန်လူမျိုးများသည် ထမင်းကို တူဖြင့် စားသော အမူအကျင့် ရှိသည်
d3	တရုတ်လူမျိုးများသည် တူ ဖြင့် စားသော အလေ့အထ ရှိကြသည်
d4	တူကို ဝါးဖြင့် ပြုလုပ်သည်
d5	တူဖြင့် မသေမချင်း ထုသတ်ခဲ့သော လူသတ်သမားကို ထောင်ချခဲ့သည်
d6	ပန်းပဲသမားသည် သံကို တူဖြင့် ထု၍ လိုရာ ပုံဖော်သည်
d7	ဖားကို တူဖြင့် ရိုက်သတ်သည်
d8	မြမြသည် ခေါက်ဆွဲကို တူဖြင့် စားလေ့ရှိသည်
d9	သံကို တူဖြင့် ရိုက်နေသည်
d10	သူ ပစ်လိုက်သော တူချောင်းသည် နံရံတွင် စိုက်သွားသည်
d11	သူမတွင် တူနှစ်ယောက် ရှိသည်
d12	သူမတို့ ညီအစ်မသည် ရုပ်ဆင်း တူသည်
d13	သူမသည် နာမည်ကျော် မင်းသမီးနှင့် ရုပ်ချင်းဆင်သည်
d14	သူမသည် အမျိုးသားကို တူဖြင့် ထုသတ်ခဲ့သည်
d15	သူသည်ခိုးဝင်လာသူကို တူဖြင့် ထုသတ်လိုက်သည်
d16	သူသည် အဖေနှင့် ရုပ်ချင်း တူသည်
d17	အမြွှာနှစ်ယောက်သည် ရုပ်ဆင်း ချွတ်စွပ်တူသည်
d18	အိမ်မြောင်ကို တူဖြင့် ထုသတ်မိသည်
d19	မောင်မောင်သည် မြမြ၏ တူတော်သည်
d20	ကြူကြူသည် အမေနှင့် ရုပ်ချင်းဆင်သည်

Figure. 2.Document Collection for Case Study

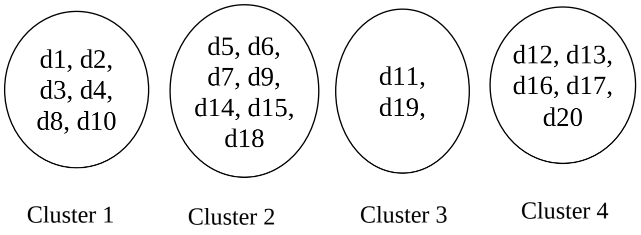


Figure. 2 Sample Clusters for 20 Documents

For the evaluation, the *F*-measure is computed as follows:
*F = 2 * ((Precision * Recall) / (Precision + Recall))*

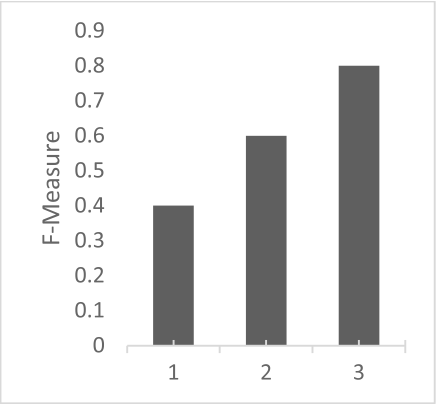
$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

The F-measure for three approaches is illustrated in Figure 5. The preliminary result shows that the proposed approach can retrieve more relevant Myanmar web documents. However, the evaluation in figure 5 is based on sample data and we plan to test with huge Myanmar document collection with large Myanmar lexicon like WordNet.

d9	သံကို တူဖြင့် ရိုက်နေသည်
d18	အိမ်မြောင်ကို တူဖြင့် ထုသတ်မိသည်
d7	ဖားကို တူဖြင့် ရိုက်သတ်သည်
d14	သူမသည် အမျိုးသားကို တူဖြင့် ထုသတ်ခဲ့သည်
d15	သူသည် ခိုးဝင်လာသူကို တူဖြင့် ထုသတ်လိုက်သည်
d5	တူဖြင့် မသေမချင်း ထုသတ်ခဲ့သော လူသတ်သမားကို ထောင်ချခဲ့သည်
d6	ပန်းပဲသမားသည် သံကို တူဖြင့် ထု၍ လိုရာ ပုံဖော်သည်

Figure. 4. Result of the query “တူဖြင့်ရိုက်”



1: Myanmar Information Retrieval with no semantic information
2: Traditional SOM Clustering
3: SOM with GloVe

Figure. 5. F-Measure of Three Approaches

In addition, the traditional recall measure has problem that concerns the impossibility of collecting exhaustive relevance judgments in a realistically large document set. Any potentially relevant document has not been missed when making relevance judgments because judges would have to go through the entire document set of nearly a million documents, which is infeasible [16]. Therefore, wewill use the pooling method to evaluate the proposed system in future. The document pool to be manually judged is constructed by putting together the top N retrieval results from a set of n systems [16].As the future work, we will evaluate and prove the result of the proposed approach by comparing traditional SOM clustering approach.

5. Conclusion

Although the semantic information retrieval for many languages have been improved, semantic information retrieval for Myanmar language still needs more researches. In this paper, we propose the novel algorithm MyanSem that is based on the parallel SOM with GloVe. This approach also modifies the distance calculation of SOM to appropriate the document association. GloVe is the word representation model that determine the words co-occurrence. This system provides the semantic information retrieval for Myanmar web pages. Therefore, the proposed system can retrieve the most relevant pages that meet the user-satisfied results. It can remove the irrelevant, meaningless pages or advertisement pages from the query result.

6. References

[1] A. Ahmad and R. Yusof, A Modified Kohonen Self Organizing Map (KSOM) Clustering for four Categorical Data, Journal of Technology (Science and Engineering), 2016, pp 75-80.

[2] B. Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", *Data-Centric Systems and Application Series, Springer-Verlag Berlin Heidelberg,* 2011.

[3] C. N. Pushpa, J. Thriveni, K. R. Venugopal and L. M. Patnaik, "Web Page Recommendation System using Self Organizing Map Technique", *International Journal of Current Engineering and Technology*, Inpressco, 2014, Vol. 4, No.5, pp. 3270-3277.

[4] C. Sadhana, L. Mary, I. Sheela, "Enhanced Self Organizing Map Algorithm for Web Usage Mining Through Neural Network", *International Journal of Trend in Research and Development*, Volume 3(6), 2016.

[5] D. Qi and C.C Li, Self Organizing Map based Web Pages Clustering using Web Logs, Proceedings of the 16th International Conference of Software Engineering and Data Engineering, SEDE, Las Vegas, Nevada, July, 2007, pp 265-270.

[6] E. Chifu and C. Cenan, "Discovering Web Document Clusters with Self –Organizing Maps", *Scientific Annals of the "Alexandru Ioan Cuza" University of Iaşi Computer Science Section*, Tome XIV, 2004, pp. 1-10.

[7] H. C. Yang, C. H. Lee, and K.-L. Ke, "TSOM: A Topic-Oriented Self-Organizing Map for Text Organization", *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering*, Vol:4, No:5, 2010.

[8] H. Yin, "The Self-Organizing Maps: Background, Theories, Extensions and Applications", *Studies in Computational Intelligence (SCI)* 115, 715–762, 2008.

[9] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation", *Empirical Methods in Natural Language Processing*, 2014, pp. 1532-1543.

[10] L. Zhang, "An Intelligent Information Retrieval Algorithm based on Knowledge Discovery and Self-Organizing Feature Map Neural Network", *International conference on Inventive Computation Technologies (ICICT)*, 2016.

[11] Martin Marcel Couturier, Disambiguating Words with Self Organizing Maps, Master Thesis, Massachusetts Institute of Technology, June, 2011.

[12] M. Negnevitsky, Artificial Intelligence, A Guide to Intelligent Systems, 2nd Edition, Pearson Education Limited 2005.

[13] M. Sasaki, "Latent Semantic Word Sense Disambiguation Using Global Co-Occurrence Information Using Non-Negative Matrix Factorization", Journal of Computer Science Applications and Information Technology, Vol (2). No. (3), pp. 1-4, 2017.

[14] N. Ampazis and S. J. Perantonis, LSISOM – A Latent Semantic Indexing Approach to Self-Organizing Maps of Document Collections, *Neural Processing Letters*, Vol 19, pp. 157-173, 2004.

[15] P. Gajdos and P. Moravec, Two-step Modified SOM for Parallel Calculation, *J. Pokorny, V. Sna sel, K. Richta (Eds.): Dateso*, 2010, pp. 13-21.

[16] S. Teufel, "An Overview Of Evaluation Methods In Trec Ad Hoc Information Retrieval And Trec Question Answering", L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 2007, pp. 163–186.

[17] T. F. Gharib, "Self Organizing Map based Document Clustering Using WordNet Ontologies", *JCSI International Journal of Computer Science Issues*, Vol. 9, Issue 1, No 2, January 2012.

[18] T. Kohonen, Self-organization and Associative Memory, *Springer-Verlag, N.Y*, 3rd edition. 1989.

[19] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela: Self-organization of a massive document collection, *IEEE Transactions on Neural Networks*, Vol. 11, no. 3, pp. 574-585, 2000.

[20] T. Munakata, Fundamentals of the New Artificial Intelligence, Neural, Evolutionary, Fuzzy and More, 2nd Edition, Springer-Verlag London Limited 2008.

[21] X. Lin, D. Soergel, G. Marchionini, "A Self-organizing Semantic Map for Information Retrieval", *Proceedings of the 14th annual international ACM SIGIR conference*, 1991, pp. 262-269.00000.00.

[22] http://learning.maxtech4u.com/self-organizing-map-som/.

Wireless and Software Defined Networking

Traffic Statistics Measurement of Video Streaming in Software Defined Network

Hla Myo Su
University of Computer Studies, Yangon
hlamyosu@ucsy.edu.mm

Aung Htein Maw
University of Information Technology
ahmaw@uit.edu.mm

Abstract

Video streaming is the dominant internet traffic. It is becoming the most popular content delivery mechanism for media services. Multiple video streams will compete for bandwidth, thus leading to degrading performance and impacting the received quality of experience (QoE). Software Defined Networking (SDN) provides a significant advantage because it is easier to introduce and tune up new functionalities. The logically-centralized control of OpenFlow brings to the operators the flexible and convenient controllability over the underlying networks. In this paper, it gets traffic statistics at each video streaming host and proposes the video streaming framework to enhance user QoE. It uses OpenFlow features to get traffic statistics from network devices. That is showing the end to end average throughput and port statistics of streamed video traffic over Mininet emulated OpenFlow networks. The experimental results of traffic statistics can be beneficially applied in the network when the network congested conditions occurred and it should be carefully interpreted considering between the clients for enhancing video streaming.

Keywords- Video Streaming, SDN, OpenFlow

1. Introduction

Although the internet has been initially designed to transfer text and data, the improvement of technologies introduced new services, such as voice and video. Nowadays, video streaming is the dominant Internet traffic. Report from Cisco Systems shows the Internet video traffic represents 59% of the global internet traffic. It will reach 77% by 2019 [7]. Therefore, it is focused on video traffic in this paper.

ISO standard: Dynamic Adaptive Streaming over HTTP (DASH), also known as MPEG-DASH, was developed by a cooperation of industries and standardized organizations aiming to high-quality video delivery. DASH runs over HTTP due to the support of HTTP by the servers, middle boxes, and client applications. Each video is available on the server with multiple copies and each copy with different encoding. Each copy is divided into chunks with equal duration. The chunk metadata is available in the Media Presentation Description (MPD). The client requests the MPD file then it chooses the most suitable bitrate and starts downloading chunks.

Traditional network architectures are rigid, it is especially hard to add new features to them. In order to overcome these issues, SDN (Software Defined Networking) paradigm has brought flexible controllability and sufficient programmability to the network operators by separating the control and data planes with an open and standardized interface. In this regard, the OpenFlow interface is the first and has become one popular protocol widely accepted. The OpenFlow standard enables the direct communication between SDN controllers and networking devices so that network management becomes easier than the traditional network management [9].

Moreover, streaming video generates the largest portion of Internet traffic, where traffic statistics measurement of video clients plays an important role in adapting to the current network load. In general, knowledge about the available bitrate of the network would benefit many users and operators of network applications and infrastructures. One of the ultimate goals in future multimedia networks is to provide a user-centric fair-share of network resources so that the user Quality of Experience (QoE) is maximized for all users in a network. There is a strong need to ensure QoE fairness across different devices in a network-wide manner [6]. To alleviate the congestions, the logically-centralized control of OpenFlow provides to the operators the flexible and convenient controllability over the underlying networks [11].

In this paper, it is interested in streaming video on the OpenFlow network. Video traffics are the majority of traffic load on the internet. These high traffic loads can potentially lead to network congestions, which are the cause of degraded quality of video at clients and also impact on QoE. The traffic statistics of the testbed are measured to solve the network congestion under the bandwidth competition of multiple clients in video streaming.

The remainder of the paper is organized as follows. Related work in this problem domain is presented in Section II. Then, Section III explains the enhance QoE video streaming framework. In Section IV, the studying experimental results in video streaming are reported. Finally, the conclusion and future work are discussed in Section V.

2. Related Work

Akhshabi et al [5] proposed traffic shaping at the server side to reduce the player oscillations. When instability in the players is detected, then the server reduces the player bitrate profile or increases it on the other case. However, this technique solves the instability problem by reducing the bitrate profile and that leads to reducing the video quality.

In [7] the authors evaluate the performance of traffic shaping for competing for video flows over a shared bottleneck link using SDN. They show the individual traffic shaping gives better results than the aggregate traffic shaping. However, the authors shape the traffic for each client to the constant value which is impractical because when the number of clients increases, then the available bandwidth will be less than the required bandwidth, which leads to poor performance.

In [2], DASH runs over HTTP which uses TCP as the transport layer, thus leading to the mismatch between the DASH adaptation logic, which runs on the client side and the TCP congestion control which runs on the server side. When two or more DASH clients compete for bandwidth, thus leads to instability of the players, unfairness between the players requested bitrates and network bandwidth under-utilization.

Abuteir et al [4] proposed NAVS (Network Assisted Video Streaming) aims to improve the QoE by reducing player instability, maximize the fairness between clients and increase the video quality. This technique can address the issue at the home network gateway without modifying the client player or the video server. The adaptation logic is the bandwidth estimation based algorithm.

To enhance video content delivery as well as increase the QoE of the end-users, [8] proposed the Server and Network Assisted DASH (SAND) architecture. SAND is a control plane for video delivery that obtains QoE metrics from the users (clients) and returns network-based measurements to help the clients enhance their overall QoE. The third-party measurement server in SAND, known as DANE (DASHassisting network element), provides measurement information to the different parties in the delivery chains including CDNs, ISPs, and content providers.

3. Proposed Effective Video Streaming Framework

The proposed framework is based on Software Defined Network architecture.

3.1. Theory Background

SDN paradigm is one of the best and most attractive solutions for enhancing the Internet with more flexibility and adaptability. It allows a logically centralized software program to control the behavior of an entire network by decoupling the routing decision tier from the forwarding layer.

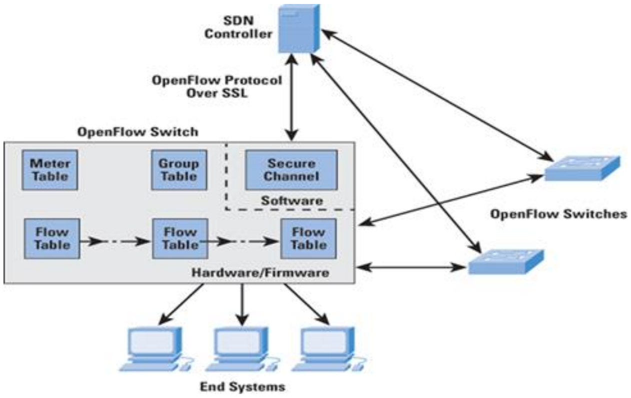


Figure 1. The Components in OpenFlow [9]

OpenFlow shown in Figure 1 enables to communicate between the control plane and data plane. When OpenFlow switch receives the new flow, the applications over control plane manipulate the first packet of this flow.

The switch interconnects with the controller and the controller directs the switch using the OpenFlow protocol. The OpenFlow switch consists of one or more flow tables, group table, and meter table. A single switch can be managed by one or more controllers. Flow tables and group table are used during lookup or a forwarding phase in order to forward the packet to the appropriate port. Meter table is used to perform simple QoS operations like rate-limiting to complex QoS operations like DiffServ. OpenFlow channel is to link to an external controller.

The controller can delete, add or update flow entries in flow tables. It makes this decision based on policies set by the administrator or depending on the conditions of the current network. OpenFlow based controllers will discover and maintain all links in the network and then will create and store all possible paths in the entire network. It can instruct switches and routers to direct the traffic by providing software-based access to flow tables. It can be used to quickly change the network layout and traffic flows as per user requirements.

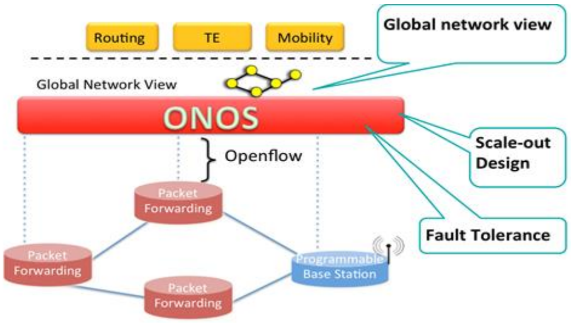


Figure 2. ONOS Controller [3]

Open Network Operating System (ONOS) controller is used in this framework. ONOS adopts a distributed architecture for high performance, availability and scalability requirements of large operator network such as

- High Throughput: up to 1M/ second
- Low Latency: 10 -100 ms event processing
- Global Network State Size: up to 1TB of data
- High Availability: 99.99% service availability

3.2. Effective Video Streaming Framework

In this framework, it is proposed to build an effective video streaming with traffic statistics measurement within the network. The functional building blocks are being brought together to form the framework. Dynamic traffic shaping approach based on the collected network traffic statistics and monitoring of video flows is proposed. It dynamically allocates bandwidth for each video flow in real time to improve performance and user QoE. In addition, it used an abstract model for the controller so its location is less important than its function. The controller consists of three functional modules as shown in Figure 3.

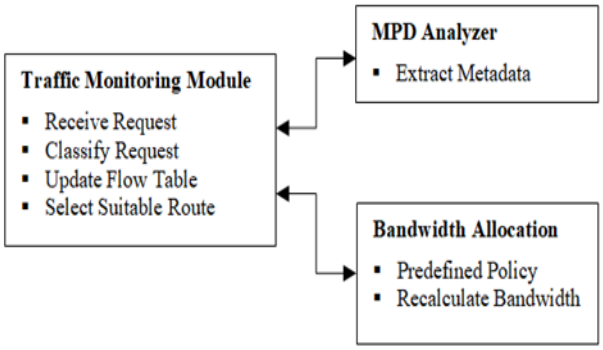


Figure 3. Functional Block Diagram of System

Multiple clients are connected to the video server for video streaming in the network. The server allows the Video-on-Demand (VoD) streaming service depends on the current network situation.

When the controller receives a request from any client, it classifies the request type, whether the video streams or not. If it is a video streaming request, the controller passes the MPD file of the client request to the MPD analyzer to start the extraction of metadata and store it in the Active Video Flow Info table. The metadata contains the video length, the number of chunks, the available bitrates and the chunks URLs. When the video flow removed from Active Video Flows Info table, the active video flow info table belonging to it is also removed. When the flow becomes inactive for a certain time that means the flow has stopped and it will be updated by removing from the active video flows table.

The bottlenecks in the network could be due to limited bandwidth in the access. When multiple video flows compete for bandwidth, a poor performance could be resulted and impact the user QoE. To improve QoE and resolve traffic congestion, it must be the bandwidth that assigned to all video flows is less than the overall bandwidth and any two video flows should get the same bandwidth. If a new video flow start or a video flow is stopped, the value of the allocated bandwidth for each client is recalculated.

4. Preliminary Experimental Results and Testbed

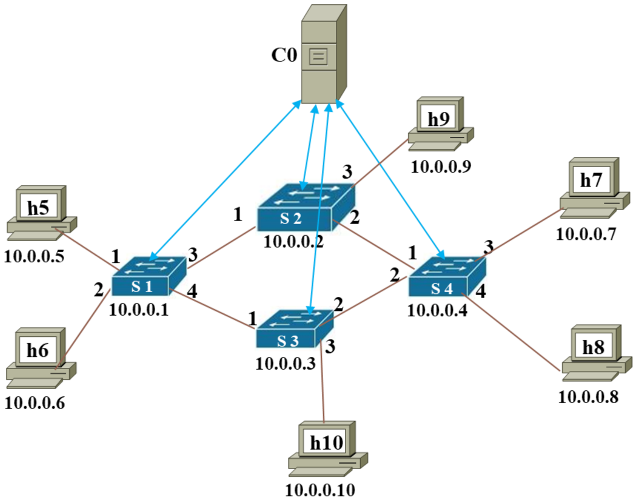


Figure 4. Testbed Topology for Emulation

In order to make an effective video streaming, the preliminary experiment on the current testbed is performed and the traffic statistics of each host are measured in this paper. To test the functionality of SDN-enabled network controls, many researchers rely on Mininet as an experimental platform. Mininet is the network emulator which can create hosts, OpenFlow switches, network links and SDN controllers virtually within a single computer. Fast prototyping can thus be achieved over Mininet platforms.

The topology of the OpenFlow network in this work is presented in Figure 4. The network consists of a video server (h5), five video clients (h6, h7, h8, h9, and h10), and four OpenFlow switches (s1, s2, s3, and s4). In addition, an SDN coordinator is equipped with an SDN controller for monitoring the video server and streaming clients. The SDN controller is based on the ONOS controller and the OpenFlow switches are based on Open vSwitch.

For the video streaming flow, on the server side, the VLC media player is used as a video streaming server. On the client side, it is also used the VLC media player as a client to receive the network video stream. On the server side of streams video, a simple HTTP (Hypertext Transfer Protocol) server bound to port 8080. There is the recommended downloading bitrate which represents the amount of bitrate required to play the selected resolution without any viewing interference. For example, YouTube requires 2.5 Mb/s for 720p and 725 Kb/s for 360p.

An animation video, Big Buck Bunny, with the two resolutions of 480x360p, 40.4 MB size and 1280x720p, 89.4MB size, and its duration of 9.56 minutes have been used for streaming over the Mininet. The video codec is H.264 with encoding bitrate. VLC media player has been configured to stream out the video packets by using TCP (Transmission Control Protocol) mode. All packets have been captured by Wireshark for all end-to-end video flows at clients.

The port data rate at video server and each video client are measured in this preliminary experiment. In Figure 5, video streaming with 360p resolution data transmission rate at every host is shown and the high definition (HD) resolution 720p video streaming testing is depicted in the next Figure 6. Host 5 transmission bit rate is the highest because of the server host.

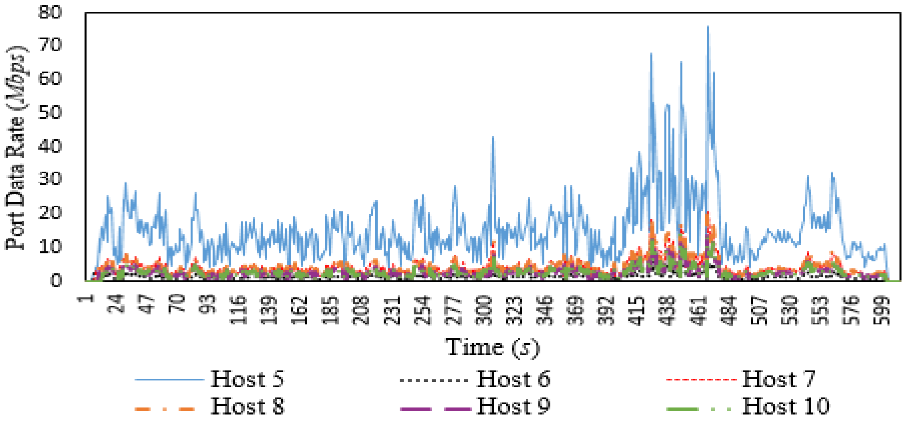


Figure 5. Data Rate Measurement for 360p Video Transmission

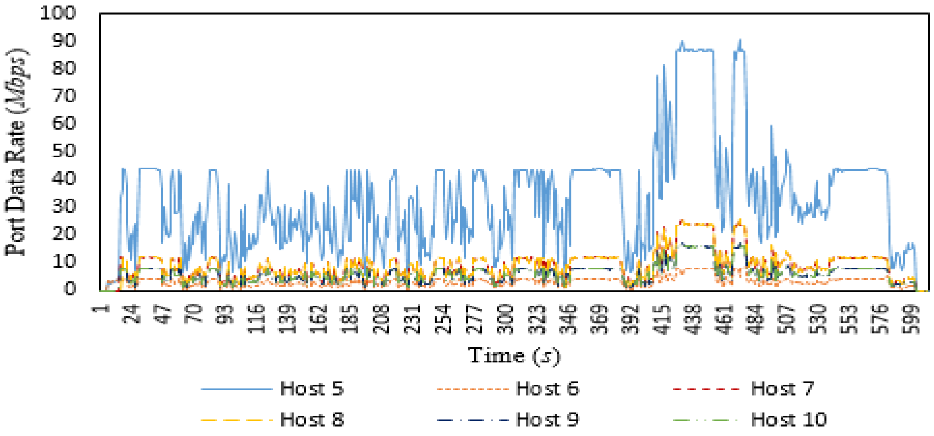


Figure 6. Data Rate Measurement for 720p Video Transmission

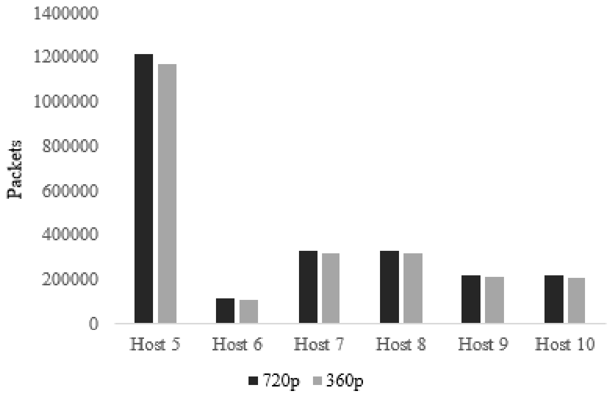


Figure 7. Video Packet Transmission with Two Resolutions

In this testbed, the HTTP video server, Host 5 transmits the video packets to all clients by using TCP mode. The number of video streaming packets from HTTP video server to all clients in TCP transmission is shown in Figure 7. Host 6 is connected to the same switch to the video server Host 5, its number of packets are lowest in the graph. The two clients Host 7 and Host 8 are attached to the same switch S4 and traversed three intermediate switches from the video server to each client, their packet amount is an approximately equal to each other and higher than the other video clients. Then, the two other hosts, Host 9 and Host 10 are streaming videos from via the next switches of server switch, so the number of packets is close to each other and medium range as their switches are located.

Table 1. Comparison of Average Bitrate

Video Streaming Hosts	360p Video Average Bitrate (*Mbps*)	720p Video Average Bitrate (*Mbps*)
Host 5	12.6322	32.5490
Host 6	1.3592	3.0058
Host 7	4.0603	8.8721
Host 8	4.0483	8.8432
Host 9	2.6989	5.9129
Host 10	2.6865	5.9149

The average transmission bitrate at each host is compared in Table 1. Their rates are different as their packet transmission amount. The nearest client needs the least rate, the video clients attached to the next hop switch of server switch are medium bitrates and the last client hosts with the highest bitrate. A link defined as a bottleneck if it provides lower data rate than the required bitrate of current streaming flow, which may cause the buffering experience at client-side. To enhance video streaming, the throughput for a client with different clients in competition for the same video stream should be fair in any situation. By maintaining the fair throughput in changing conditions, the more effective video streaming is led to enhance user experience.

5. Conclusion

In this paper, it is reported the traffic statistics of the current testbed such as the data transmission rate, number of packets and average bitrate at each host in video streaming with two different resolutions. This preliminary results can assist to implement the proposed video streaming framework used ONOS SDN. The proposed framework tends to allocate bandwidth for different video streams based on the predefined policy, the number of video flows, available video bitrates, and bottleneck bandwidth. If a new video flow starts or existing video flow is stopped, the policy is easily modified and the value of the allocated bandwidth for each client will be recalculated. The maximize bitrate received by each client can be reduced the bottleneck network traffic even busy hour duration with higher resolution. The current testbed topology in emulation consists of only one video server for simplicity of preliminary test. This is not covered in the real world. The implementations of the proposed framework by adding more video servers is in future work.

6. References

[1] CISCO, “Visual networking index: Global IP traffic forecast 2014 - 2019”, May 2015.

[2] V. Anthony and S. Iraj, “The mpeg-dash standard for multimedia streaming over the internet,” *IEEE Multimedia*, No. 4, 2011, pp. 62-67.

[3] P. Berde, M. Gerola, J. Hart, Y. Higuchi, M. Kobayashi, T. Koide, B. Lantz, B. O. Connor, P. Radoslavov, W. Snow and G. Parulkar, “ ONOS, Towards an Open, Distributed SDN OS”, *Proceedings of the third workshop on Hot topics in software defined networking,* Chicago, USA, August 22, 2014, pp. 1-6.

[4] R. M. Abuteir, A. Fladenmuller, and O. Fourmanx, “SDN based architecture to improve streaming in home network”, *IEEE 30th International Conference on Advanced Information Networking and Applications*, 2016, pp. 220-226.

[5] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen, “Server-based traffic shaping for stabilizing oscillating adaptive streaming players,” in *Proceeding of the 23rd ACM Workshop on Network and Operating Systems Support for Digital Audio and Video,* ACM, 2013, pp. 19-24.

[6] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, “Towards network-wide QoE fairness using openflow-assisted adaptive video streaming,” in *Proceedings of the 2013 ACM SIGCOMM workshop on Future human-centric multimedia networking*, ACM, 2013, pp. 15-20.

[7] J. J. Quinlan, A. H. Zahran, K. Ramakrishnan, and C. J. Sreenan, “Delivery of adaptive bit rate video: balancing fairness, efficiency and quality,” in *Local and Metropolitan Area Networks (LANMAN), 2015 IEEE International Workshop* on IEEE, 2015, pp. 1-6.

[8] *ISO/IEC JTC1/SC29/WG11 (MPEG) Report, Technical report*, ISO, November 2013.

[9] W. Stallings, “Software-Defined Networks and OpenFlow”, *The Internet Protocol Journal*, Volume 16, No.1, http://www.cisco.com/.

[10] M. M. Rahman, “Introduction to SDN”, ppt, November 2013, moin@1asia-ah1.com, http://slideshare.net.

[11] P. Panwaree, J. W. Kim and C. Aswakul, “Packet Delay and Loss Performance of Streaming Video over Emulated and Real OpenFlow Networks”, *29th International Technical Conference on Circuit/Systems, Computers and Communications (ITC-CSCC)*, Phuket, Thailand, July 1-4, 2014, pp. 777- 779.

Applications of Minimax Access-Point Setup Optimization Approach to IEEE802.11ac WLAN at 5GHz

Kyaw Soe Lwin*, Zinan Wang**, Nobuo Funabiki**, MinoruKuribayashi**,
and Wen-Chung Kao***
**West Yangon Technological University, Myanmar*
***Graduate School of Natural Science and Technology, Okayama University, Japan*
****Department of Electrical Engineering, National Taiwan Normal University, Taiwan*
kyawsoelwin@ytu.edu.mm ** pr3y3hgk@s.okayama-u.ac.jp{funabiki, kminoru}@okayama-u.ac.jp***jungkao@ntnu.edu.tw*

Abstract

Previously, we have studied the minimax access-point (AP) setup optimization approach to improve the throughput performance of a wireless local-area network (WLAN) with IEEE 802.11n at 2.4GHz. Recently, IEEE802.11ac at 5GHz has become popular due to the much higher data rate, using a larger number of antennas for multiple-input-multiple-output (MIMO), the larger frame aggregation size, the beamforming, and the multi-user-MIMO(MU-MIMO). In this paper, we present the application of the minimax AP setup optimization approach to WLAN for 11ac at 5GHz. First, the throughput performance of this WLAN using MIMO links is investigated and compared with the one for 11n at 2.4GHz. Then, the minimax approach is applied with slight modifications. The effectiveness of our approach for 11ac at 5GHz is confirmed through extensive experiments in three network fields.

Keywords-Wireless local-area network, access-point setup, MIMO, IEEE 802.11n/ac, throughput estimation model

1. Introduction

Nowadays, *IEEE 802.11 wireless local-area network (WLAN)* has become common in daily life as the high-speed and cost-effective Internet access network. Previously, we have studied the *minimax AP setup optimization approach* to improve the throughput performance of WLAN [1]. In this approach, the *bottleneck host*, which receives the weakest signal from the AP in the field, is detected using the *throughput estimation model*. Then, the AP setup is optimized by changing the height, orientation, and coordinate, such that the throughput of this bottleneck host is maximized.

This throughput estimation model consists of two functions. First, the *received signal strength*(RSS) at the receiver is estimated by the *log-distance path loss model* [2]. Second, this RSS is converted to the throughput using the *sigmoid function*. The parameter values of these functions are optimized using the *parameter optimization tool* with the throughput measurement results at the WLAN.

However, in this previous study, we only considered WLAN with *IEEE 802.11n at 2.4GHz* although *IEEE 802.11ac at5GHz*has become popular due to the much higher data rate than 11n, using a larger number of antennas for *multiple-input-multiple output (MIMO)*, the larger *frame aggregation* size, the *beamforming*, and the *multi-user-MIMO (MUMIMO)* [3]. For example, the maximum throughput of a commercial AP *NEC WG2600HP* can be 1,733*Mbps* for 11acat 5GHz and 800*Mbps*for 11n at 2.4GHz [4].

In this paper, we present the application of the minimax AP setup optimization approach to WLAN with IEEE 802.11ac at 5GHz. First, the throughput performance of this WLAN using MIMO links is investigated and compared with the one for 11n at 2.4GHz. Then, the minimax approach is applied to 11ac at 5GHz with slight modifications. The effectiveness of our approach for 11ac at 5GHz is confirmed through extensive experiments in three network fields.

The rest of this paper is organized as follows: Section 2. presents the related works to this paper. Section 3. reviews the previous minimax AP setup optimization approach for 11n at 2.4GHz. Section 4. introduces the application of the minimax approach to 11ac at 5GHz. Finally, Section 5. concludes this paper with future works.

2. Related Works

Several related works have been reported in literature. In [5], Kriara et al. studied the performance characterization of IEEE 802.11ac in terms of the throughput, the jitter, and the fairness using real testbed deployments. The authors reported that 11ac WLAN with wider channels can be fairer in dense environments with high interferences.

In [6], Simić et al. studied the combined impacts of the channel bandwidth, the traffic profile, and the AP density and placement on the overall network-level throughput and the fairness in IEEE 802.11ac. The authors evaluated the performance of 11ac Wi-Fi in a large 24-node indoor testbed. They addressed that wide 80MHz channels are only beneficial in very dense deployments with extreme offered traffic volumes, due to the significant adjacent channel interference (ACI) which couples narrower channels.

In [7], Newell et al. carried out the performance evaluations of IEEE 802.11n and 11ac to characterize the effects of the distance and the interference between different channels. The authors concluded that throughput performance of 11ac decreases at an extremely faster rate with the increasing distance from the client to the AP when compared to 11n at 5GHz.

A variety of studies have been addressed to the radio wave propagation regarding positions, polarizations, and radiation patterns of transmitting and receiving antennas [8]-[11]. It is revealed that different antenna configurations and orientations have significant impacts on performances of MIMO links.

3. Review of Minimax AP Setup Optimization Approach for 11n at 2.4GHz

In this section, we review the minimax AP setup optimization approach for IEEE 802.11n at 2.4GHz.

3.1. Overview

In the minimax AP setup optimization approach, first, the throughput is measured for each link between the target AP and a possible host location in the network field. Next, the parameters of the *throughput estimation model* are tuned based on the throughput results using the *parameter optimization tool*. Then, the *bottleneck host* in the field is found through simulations using the model. After that, the AP setup is optimized by changing the height, orientation, and coordinate such that the throughput of this bottleneck host is maximized against the AP. Finally, we evaluate the overall throughput improvement among the hosts by the AP setup optimization.

3.2. Throughput Estimation Model

The throughput estimation model has been developed to accurately estimate the throughput of a wireless communication link between an AP and a host in WLAN from the network field information. First, this model estimates the RSS at the host using the *log-distance path loss model*, which considers the distance and the obstacles between AP and the host. Next, it converts the RSS to the throughput using the *sigmoid function*. Both functions have several parameters whose values can affect the estimation accuracy.

3.2.1. Signal Strength Estimation. The *RSS* at a host from an AP is calculated using the *log-distance path loss model* [2]:

$$P_d = P_1 - 10\,\alpha\log_{10} d - \sum_k n_k W_k \qquad (1)$$

where $P_d$represents the RSS (*dBm*) at the host, αdoes the path loss exponent factor, d does the distance (m) to the host from the AP, P_1 does the RSS (*dBm*) at the host at the $1m$ distance from the AP when no obstacle exists between them, n_k does the number of $type\ k$ obstacles along the path between the AP and the host, and$W_k$does the signal attenuation factor (*dB*) for the $type\ k$ obstacle. P_1 , α ,and W_k are parameters to be tuned. To consider the multipath effect, the *indirect path* is also considered by selecting a *diffraction point* for each AP/host pair and select the larger *RSS* between the direct and indirect signals for sigmoid function. It is noted that α can be replaced by α_{inc} (enhanced path loss exponent factor) for$d \geq d_{thr}$ (distance threshold) to improve the estimation accuracy [13].

3.2.2. Throughput Conversion. The *RSS* is converted to the throughput or the data transmission speed from the AP to the host using the *sigmoid function*:

$$\mathrm{S} = \frac{a}{1+\exp(-\frac{(120+P_d)-b}{c})} \quad (2)$$

where Srepresents the estimated throughput (Mbps) when the RSS (dBm) at the host is P_d. a,b, and $c$are parameters to be tuned.

3.3. Parameter Optimization Tool

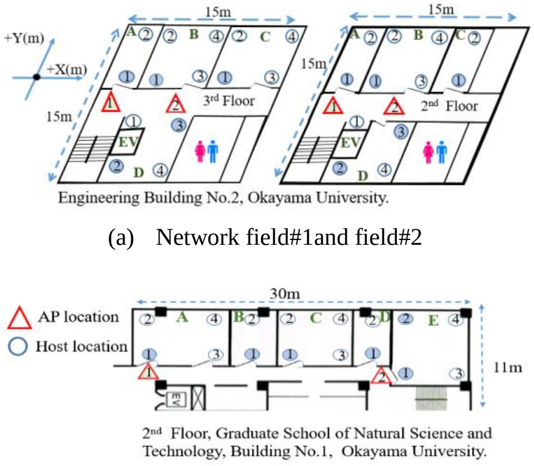
The throughput estimation model has a number of parameters whose value determines the estimation accuracy. These values are optimized by use of the *parameter optimization tool*, which adopts a local search algorithm that combines the tabu table and the hill climbing procedure to avoid a local minimum. This tool has normally been implemented in the general form, so that it can be used for a variety of algorithms/logics that have parameters to be optimized. The program for the tool has been independently implemented from the program with the throughput estimation model. It runs the model program as its child process. The optimality of the current parameter values in the model program is evaluated by the throughput estimation error that is given in the output file.

3.4. Evaluation Results

In this section, we present the evaluation results of the minimax approach for 11n at 2.4 GHz in three network fields.

3.4.1. Network Fields and Devices. The outdoor network field in Figure 1 and three network fields in Figure 2 are considered. In each indoor field, the triangle represents the AP location and the circle does the host location. One NEC WG2600HP with four internal antennas is used for the AP. Two laptop PCs with Windows OS for the server and the client host, where the host PC has dual-band Wireless-AC 8260 wireless adapter. Two-stream IEEE 11n MIMO links with the 40MHz channel at 2.4GHz are generated in measurements. *iperf* [12] is used for throughput measurements by generating TCP traffics for 50*sec* with 477*Kbytes* window size and 8*Kbyte* buffer size. It is noted that all the experiments were conducted on weekends to reduce the interferences from other wireless devices and human movements.

Figure 1. Outdoor network field


15m
15m
+Y(m)
+X(m)
15m
15m
A
B
C
D
EV
3rd Floor
2nd Floor
Engineering Building No.2, Okayama University.
30m
AP location
Host location
A
B
C
D
E
EV
11m
2nd Floor, Graduate School of Natural Science and Technology, Building No.1, Okayama University.

(a) Network field#1and field#2

(b) Network field#3

Figure 2. Three indoor network fields

3.4.2. Throughput Estimation Results. For the outdoor network field, Figure 3 shows the throughput measurement results when the distance between the AP and the host is changed from 0m to 170m with the interval of 5m. The estimated throughput by the model is also illustrated there.

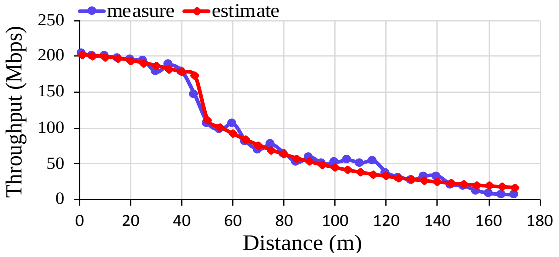


Figure 3. Throughput results in outdoor network field

For the three indoor network fields, we follow the throughput measurement minimization procedure in [13] to reduce the labor cost of throughput measurements. It selects the limited host locations for throughput measurements to optimize the parameter values while keeping the accuracy. Then, five host locations are selected in field#1 and field#2, and six host locations are in field#3. For example, the shaded circle in Figure 2 indicates the selected host location for AP1 respectively.

After throughput measurements, the parameter values of the throughput estimation model were optimized by applying the tool to the results in each field. Table 1 displays the parameter optimization results for 11n at 2.4GHz. The parameter optimization tool [1] is applied to the measurement data in each field. To verify the accuracy of the model, the throughput estimation errors (Mbps), given by the difference between the measured throughputs and the estimated ones, are calculated.

Table 1. Parameter optimization results for 11n at 2.4GHz

Parameter	Out.	field#1	field#2	field#3
P_1	-20	-35.6	-35.2	-34
α	2.4	2.09	2.20	2.10
α_{inc}	2.9	2.19	2.40	2.20
d_{thr}	45	5	5.2	5
corridor wall	~	7	7	7
partition wall	~	8	8	5
intervention wall	~	7	7	~
glass wall	~	~	2	~
elevator wall	~	2	2.8	~
Door	~	3	3	2
diffraction point	~	2	2	1.9
a	202	190	190	194
b	49.5	46.5	50	40
c	6	7	8	6.5

Table 2 summarizes the average, the maximum, the minimum, the standard deviation (SD), and the coefficient of variation (CV) of the estimation errors for all the links between the host locations and each AP. Table 2 indicates the high accuracy of the model. It is noted that the bottleneck host providing the lowest throughput by the model is coincident with the one found by the measurements for any A Plocation. Specifically, in field#1, C4 is the bottleneck host for AP1 and A2 is for AP2. In field#2, C2 is for AP1 andA2 is for AP2. In field#3, D2 is for AP1 and A2 is for AP2.

Table 2. Throughput estimation errors (Mbps) for 11n at2.4GHz

field	AP	Mea.	Estimation errors				
		Avg. TP	avg.	max.	min.	SD	CV (%)
Out.	AP	88.34	6.35	27.44	0.27	5.87	6.64
#1	1	139.63	11.83	36.58	2.11	8.65	6.19
	2	155.36	8.61	24.82	0.39	6.94	4.47
#2	1	141.28	15.78	24.64	3.98	6.34	4.49
	2	166.00	12.29	20.46	2.58	6.65	4.01
#3	1	147.17	12.18	20.32	0.01	5.59	3.80
	2	161.56	9.69	24.80	0.24	7.54	4.67

3.4.3. AP Setup Optimization Results. In the three indoor network fields, the setup condition for each AP is optimized to maximize the throughput of the bottleneck host. To evaluate the effectiveness of the optimization, the average throughput improvement of all the hosts is investigated for each AP. The average throughputs of three cases, 1) the original setup, 2) after the height and orientation optimizations (after H&O), and 3) after all the optimizations (after ALL), are compared, where the improvement rates from 1) to 2), and those from2) to 3) are also calculated.

Table 3 reveals the results which indicate that the height and orientation optimization can improve the average throughput for any AP, while the coordinate shift does not improve it for specific APs, because the multipath effect is not sensitive to the link environment for MIMO.

Table 3. Average throughput improvements of 11n at2.4GHz

field	AP	1)orig. setup (Mbps)	2) after H&O (Mbps)	imp. Rate From 1) (%)	3)after ALL (Mbps)	imp. Rate From 2) (%)
#1	1	139.63	145.94	4.51	145.94	0.00
	2	155.36	179.86	15.77	182.00	1.19
#2	1	141.28	155.59	10.13	160.33	3.05
	2	166.00	174.92	5.37	174.92	0.00
#3	1	147.17	156.38	6.26	156.38	0.00
	2	161.56	169.43	4.87	171.00	0.93

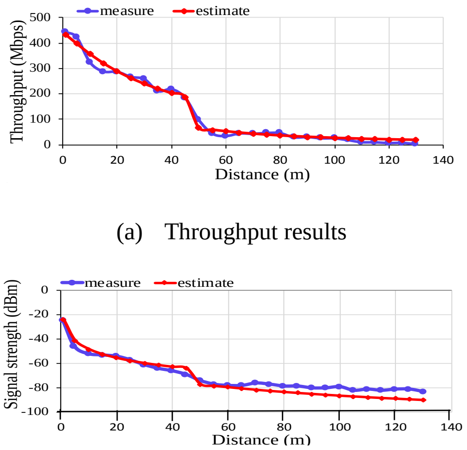
4. Minimax AP Setup Optimization Approach to 11ac at 5GHz

In this section, we present the application of the minimax AP setup optimization approach to WLAN with IEEE 802.11ac at 5GHz.

4.1. Throughput Measurement Results

To investigate the performance of IEEE 802.11ac at 5GHz with the 80MHz channel, firstly we conduct the measurements in outdoor and indoor network fields using the same devices and software tools as in Section 3. 4. 1. It is noted that the wireless devices support 11ac. That is, the 80MHz channel is used as the default setting for this commercial AP.

4.1.1. Outdoor Network Field. Figure 4 exhibits the outdoor field measurement and estimated results for the throughput and RSS at the host. It is observed that they sharply drop at around 50m distance, and the throughput is lower than that of 11n at 2.4GHz. Besides, the signal at the host is lost at 140m or larger.



(a) Throughput results

(b) RSS results

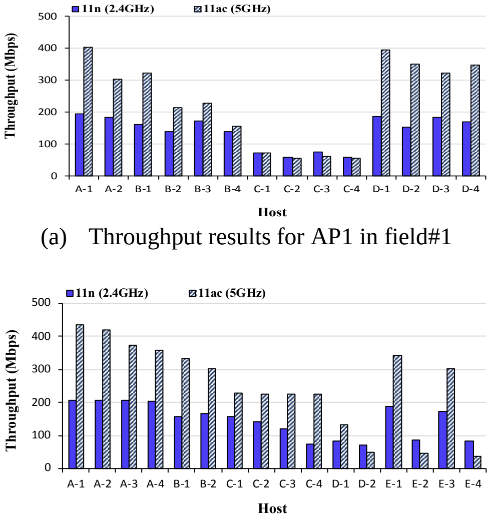
Figure 4. Measurement results in outdoor network field

4.1.2. Indoor Network Fields. Figure 5 shows indoor measured throughput results for AP1 in field#1 and in field#3 in Figure 2. These results demonstrate that when the host is near the AP such as A-1 in both fields, the throughput by 11ac becomes more than double of that by 11n due to the wider channel bandwidth. However, as the distance between the host and the AP becomes larger, the throughput advantage of 11ac will turn out to be smaller due to the higher frequency. Furthermore, at certain host locations such as C-2, C-3 in field#1 and D-2, E-2, E-4 in field#3 where several walls exist along the line-of-sight from the AP, the throughput for 11ac is smaller than that for 11n.

4.2. Throughput Estimation Model

Next, we present the throughput estimation model for the IEEE 802.11ac link at 5GHz.

4.2.1. Exclusion of Slow Host Locations. With the IEEE 802.11ac at 5GHz, the *log-distance path loss model* in the throughput estimation model may not be accurate for a slow link that has a small throughput, because the RSS at the receiver becomes too small due to the larger path loss at the higher frequency. Besides, the throughput range of the 11ac link at 5GHz can be much larger than the11n link at 2.4GHz.



(a) Throughput results for AP1 in field#1

(b) Throughput results for AP1 in field#3

Figure 5. Throughput measurement results for two APs in indoor environments

Therefore, in this paper, any host location whose throughput is smaller than 100*Mbps*, where the big drop of the throughput is observed in Figure 4 (a), is excluded from the scope of the throughput estimation model and the bottleneck host selection in the AP setup optimization. For instance, C-1, C-2,C-3, C-4 for AP1 in field#1 and D-2, E-2, E-4 for AP1 infield#3 are excluded. It is noted that $100Mbps$ is selected from the results in Figure 5.

4.2.2. Model Parameter Optimization Results. The parameters of the throughput estimation model for 11ac at 5GHz are optimized by using the parameter optimization tool with the remaining throughput measurement results in each field. Table 4 shows their values. It is observed that α (path loss exponent factor) becomes larger to α_{inc} at shorter distance threshold $d \geq d_{thr}$ than 11n at 2.4GHz, because of the larger path loss for 11ac at 5GHz. The value of a in the sigmoid function becomes more than double due to the larger throughput range.

Table 4. Parameter optimization results for 11ac at 5GHz

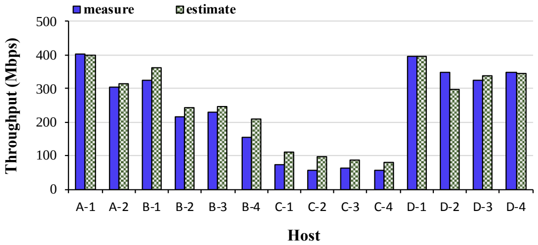
Parameter	Out.	field#1	field#2	field#3
P_1	-24	-35.1	-34	-34
α	2.4	2.00	2.00	2.00
α_{inc}	3.1	2.50	2.50	2.50
d_{thr}	45	5	4	5
corridor wall	~	8	8	7
partition wall	~	8	8	6
intervention wall	~	7	6	~
glass wall	~	~	2	~
elevator wall	~	2	2	~
Door	~	2.6	2.5	3
diffraction point	~	2	1	1
A	442	445	437	452
B	59	53.5	51.85	42.0
C	9	9	8	9

4.2.3. Throughput Estimation Results. To verify the accuracy of the throughput estimation model, the estimated throughput results are compared with the measurement results. Table 5 summarizes the average, the maximum, the minimum, the standard deviation (SD), and the coefficient of variation (CV) of the throughput estimation errors (Mbps) for each AP. The CV is similar between 11ac at 5GHz and 11n at 2.4GHz for most of the APs. Thus, similar estimation accuracy is maintained for 11ac at 5GHz. Figure 6 shows the measured and estimated throughput results for AP1 in field#1 and field#3. It indicates that the estimation error for the slow host whose throughput is smaller than 100*Mbps* is large.

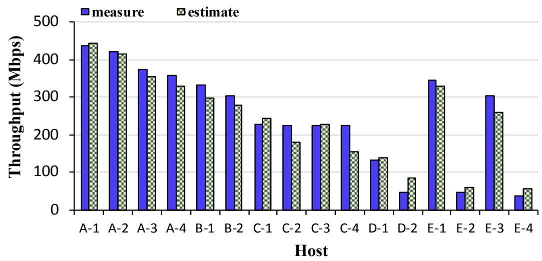
Besides, it is found that for any AP, the bottleneck host found by the model is coincident with the one of the measurements. Specifically, in field#1, B-4 is the bottleneck host for AP1, and A-2 is for AP2. In field#2, C-2 is for AP1, and A-2 is for AP2. In field#3, D-1 is for AP1, and A-2 is for AP2. These results justify the use of the throughput estimation model in the minimax AP setup optimization approach for 11ac at 5GHz.

Table 5. Throughput estimation errors (Mbps) for 11ac at5GHz

field	AP	Mea.	Estimation errors				
		Avg. TP	avg.	max.	min.	SD	CV (%)
Out.	AP	126	11.90	35.64	0.13	9.81	7.79
#1	1	305.20	22.49	53.91	1.88	20.04	6.57
	2	305.43	25.26	66.49	0.02	17.81	5.83
#2	1	298.42	24.91	66.68	2.82	16.31	5.47
	2	319.92	19.63	41.90	4.12	12.16	3.80
#3	1	302.85	23.57	69.73	2.15	20.51	6.77
	2	349.25	33.64	62.33	5.00	15.69	4.49



(a) Throughput results for AP1 in field#1



(b) Throughput results for AP1 in field#3

Figure 6.Measured and estimated throughput results for 11ac at 5GHz

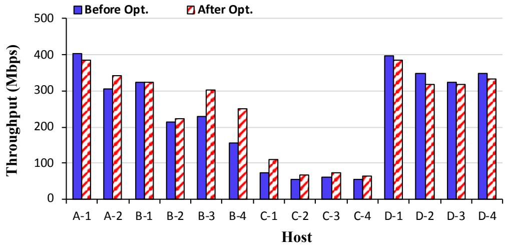
4.3. AP Setup Optimization Results

Finally, we apply the minimax AP setup optimization approach to IEEE 802.11ac at 5GHz in the three indoor net-work fields. Table 6 shows the results of each of the six APs in the three network fields. This table indicates that our approach can improve the average throughput in any field. To elaborate, for AP2 in field#2, the average through- put is improved from 319.92Mbps to 350Mbps, which means 9.4% improvement. Thus, the effectiveness of the minimax approach for IEEE 802.11ac at 5GHz is confirmed.

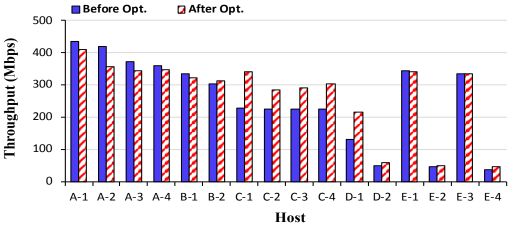
Figure 7 compares the measured throughput results for AP1 in field#1 and field#3 before and after applying the optimal approach. It can be noticed that throughputs after optimization become more averaged among the host locations than those before optimization.

Table 6. Average throughput improvements of 11ac at5GHz

field	AP	1)orig. setup (Mbps)	2) after H&O (Mbps)	imp. Rate From 1) (%)	3)after ALL (Mbps)	imp. Rate From 2) (%)
#1	1	305.20	315.40	3.34	318.60	1.01
	2	305.43	317.29	3.88	317.29	0.00
#2	1	298.42	310.33	3.99	317.75	2.39
	2	319.92	350.00	9.40	350.00	0.00
#3	1	302.85	317.38	4.80	323.31	1.87
	2	349.25	354.13	1.40	354.13	0.00



(a) Throughput improvement results for AP1 in field#1



(b) Throughput improvement results for AP1 in field#3

Figure 7. Throughput improvements by setup optimizations for 11ac at 5GHz

5. Conclusion

In this paper, we presented the application of the minimax AP setup optimization approach to WLAN for IEEE 802.11ac at 5GHz. First, the throughput performance of this WLAN using MIMO links was investigated and compared with the one for 11n at 2.4GHz. Then, the minimax approach is applied with slight modifications, where the effectiveness is confirmed through extensive experiments. In future works, we will apply this approach to various network fields.

6. References

[1] K. S. Lwin, N. Funabiki, C. Taniguchi, K. K. Zaw, M. S. A. Mamun, M. Kuribayashi, and W.-C. Kao, “A minimax approach for access point setup optimization in IEEE 802.11n wireless networks”, Int. J. Netw. Comput., vol. 7, no. 2, pp. 187-207, July 2017.

[2] D. B. Faria, “Modeling signal attenuation in IEEE 802.11 wireless LANs”, Tech. Report, TR-KP06-0118, Stanford Univ., July 2005.

[3] M. S. Gast, 802.11ac: a survival guide, 1st ed., O’Reilly, 2013.

[4] NEC Inc., “WG2600HP manual”, http://www.aterm.jp/support/manual/pdf/am1-002673.pdf.

[5] L. Kriara, E.C. Molero, and T.R. Gross, “Evaluating 802.11ac features in indoor WLAN: an empirical study of performance and fairness”, Proc. ACM Int. Work. Wireless Netw. Testb. Exp. Eval. Char., pp. 17-24, 2016.

[6] L. Simić, J. Riihijärvi, and P. Mähönen, “Measurement study of IEEE 802.11ac Wi-Fi performance in high density indoor deployments: are wider channels always better? ”, Proc. IEEE Int. Symp. World. Wireless, Mobil. Multi. Netw. (WoWMoM), pp. 1-9, 2017.

[7] D. Newell, P. Davies, R. Wade, P. Decaux, and M. Shama, “Comparison of theoretical and practical performances with 802.11n and 802.11ac wireless networking”, Proc. IEEE Int. Conf. Adv. Inform. Netw. Appl. Work. (WAINA), pp. 710- 715, 2017.

[8] Z. Sun and I. Akyildiz, “Channel modeling and analysis for wireless networks in underground mines and road tunnels”, IEEE Trans. Commun., vol. 58, no. 6, pp. 1758-1768, 2010.

[9] D. Kong, E. Mellios, D. Halls, A. Nix, and G. Hilton, “Throughput sensitivity to antenna pattern and orientation in 802.11 n networks”, Proc. IEEE Int. Symp. Person. Indoor. Mobil. Radio Commun. (PIMRC), pp. 809-813, 2011.

[10] P. Y. Qin, Y. J. Guo, and C. H. Liang, “Effect of antenna polarization diversity on MIMO system capacity”, Proc. IEEE Antenna. Wireless Propa., vol. 9, pp.1092-1095, 2010.

[11] H. C. Lo, D. B. Lin, T. C. Yang, and H. J. Li, “Effect of polarizationon the correlation and capacity of indoor MIMOchannels”, Int. J. Antenna. Propa., 2012.

[12] ACD.net, Iperf Speed Testing, http://support.acd.net/wiki/index.php?title=IperfSpeedTesting.

[13] K. S. Lwin, K. K. Zaw, and N. Funabiki, “Throughput measurement minimization for parameter optimization of throughput estimation model”, Proc. Chugoku-Branch J.Conf., Oct. 2017.

[14] S. K. Debnath, M. Saha, N. Funabiki, and W.-C. Kao, “A throughput estimation model for IEEE 802.11n MIMO link in wireless local-area networks”, Proc. Int. Conf. Comput. Commun. Syst. (ICCCS), pp. 327-331, April 2018.

Evaluation of QoS Provisioning over Software Defined Network using Segment Routing

Ohmmar Min Mon, Myat Thida Mon
University of Information Technology, Yangon, Myanmar
ommm@uit.edu.mm,myattmon@uit.edu.mm

Abstract

Provisioning quality of service (QoS) is a big deal to deliver different applications over the current internet. With the advancements of using multimedia applications, the necessity of Quality of Service (QoS) is increasing rapidly. As real-time applications increase, Software Defined Network (SDN) has emerged as a well-established paradigm for next generations networks. By utilizing the characteristics of SDN, this paper proposes QoS provisioning based segment routing (SR) over SDN framework to find the feasible path according to the QoS requirements. This QoS provisioning architecture includes monitoring of link states among switches and providing of flow's QoS requirements. This QoS provisioning is the use of the available bandwidth to react to the network traffic. The routing algorithm solves the problem of inefficient bandwidth. If there is no available bandwidth path, the controller will be decided depending on the flows by using the proposed algorithm. Simulation results are presented to show the effectiveness of QoS provisioning using OpenFlow/ONOS controller over SDN environment.

Keywords- Software Defined Network, QoS provisioning, Segment routing, OpenFlow

1. Introduction

Software Defined Network (SDN) has emerged as a new paradigm that can be implemented to adapt the existing network function. Providing QoS guarantee can give a strong guarantee to end hosts. With the development of the Internet, as a larger-scale networking system faces some unexpected challenges to satisfy various services request. Some applications, such as Voice over IP (VoIP), multimedia, video conferencing, HDTV etc. have been getting increasingly popular on the Internet. To meet the demand for QoS requirements, there is a Service Level Agreement (SLA) [1] between business customers and a service provider. There are many QoS parameters such as delay, bandwidth, jitter, loss probability, and cost, but the important one is bandwidth. If bandwidth for a packet flow is not enough, congestion will occur in bottleneck links, which causes severe packet drops and increases end-to-end delay.

SDN simplifies the QoS routing process and evolves rapidly. It has more advanced features while using traditional network function. SDN Controller receives the information from all switches in the network and based on the received information [9] as well as available network bandwidth information, a controller can build the network topology. Traditional network monitoring techniques such as NetFlow and sFlow support various kinds of measurement tasks. OpenFlow managed by the Open Networking Foundation (ONF) is the first popular implementation of SDN. The OpenFlow Switch performs data forwarding process based on the decision made by the Controller. Segment Routing (SR) is a new emerging traffic engineering technique and SR header contains a sequence of segments Segment Identifiers (SIDs) [12]. The segment labels are carried in the packet header and so per-flow state is maintained only at the ingress node. SR controller can take advantage of the possible segment routing by choosing segments based on the traffic. Signaling protocols are not required to accomplish resource reservation. The main challenge is how to achieve the best path for QoS flow. This paper takes full advantage of SDN's characteristic to implement QoS framework. This paper implements QoS provisioning for the available bandwidth for each application flow. The goal is to enable QoS provisioning in OpenFlow as one implementation of ONOS SDN.

The remainder of the paper is structured as follows: Section II briefly reviews the related work. Section III outlines SDN and segment routing architecture. QoS provisioning in SDN is proposed in Section IV of this paper. The performance with the evaluation experiments and test results on SDN testbed is discussed in Section V. Finally, section VI gives the conclusions and our future research.

2. Related Work

Several QoS routing algorithms have been suggested to achieve the best path using QoS aware routing algorithms. QoS problem with bandwidth and delay requirements using simulated annealing based QoS-aware routing (SAQR) algorithm to find the best fit path is solved in [1]. However, it considered L2 legacy switches with SDN switches. The performance of the network by separating the application into bandwidth oriented and latency oriented application using routing algorithms such as Maximum Delay-Weighted Capacity Routing Algorithm (MDWCRA), Minimum Interference Routing Algorithm (MIRA) and Dynamic Online Routing Algorithm (DORA) for the multi-domain network is presented to increase network capacity in [3]. This system discussed that number of SD pairs affect BRR performance of the considered different routing algorithms.

QoS routing methods depending on application requirements and link cost values to measure the maximum bandwidth and delay between the proposed algorithm and traditional shortest path algorithm using Dijkstra algorithm are described in [4]-[5].

Available bandwidth is an important dynamic characteristic of a network path. Here [6] used the passive method to measure available bandwidth for any time. They discussed the bandwidth measurement overhead due to the passive way and [7] solved the problem of the lack of timestamp using OpenFlow. In our approach, we used the results obtained in different network configurations. The adaptive video is video streaming and DASH [8] is expected to be the future standard for adaptive video transfer. It discussed to obtain the appropriate path for video flows depending on the segment. It also considers the available bandwidth and bitrate of the segment.

R.Kumar [10] proposed mechanisms that provide end-to-end delays for critical traffic in real time systems using COTS SDN switches. This system shows that increasing the number of flows slightly decreases end-to-end delays. And [11] presented SDN/OpenFlow control framework that provides bandwidth guarantees for priority flows and implemented the experiments that proved its benefits in comparison with best-effort shortest path routing and IntServ. Our approach used the results for the available bandwidth of QoS flows using segment routing.

3. Software Defined Network and Segment Routing

In the SDN architecture, network architecture and the network intelligence is separated from the data plane. Forwarding is handled as flows. The controller has a logically centralized view of the flow, removing the requirement to carry such administrative information in packets. SDN scene has experienced significant growth in the number of projects and is investigated for various network functionalities such as security, quality of service etc. The most used standard is OpenFlow as shown in Figure 1. From a scale and simplicity perspective, Segment Routing is especially powerful in the era of SDN with application requirements programming the network behavior. SDN controller intelligence is used to map the optimal path onto segments. Segment Routing enables to use non-shortest paths by specifying alternative routes. Packets are forwarded through the shortest path from the source to the first segment, then to the second segment and so on.

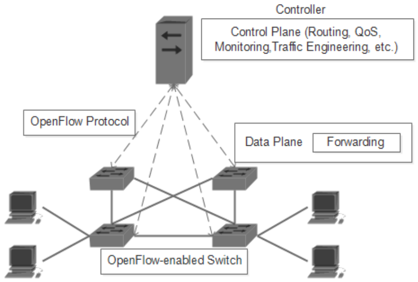
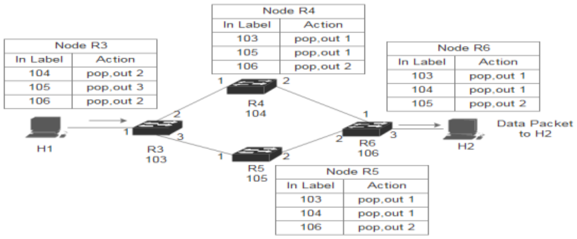


Figure 1. Software Defined Network Infrastructure

There are three actions that are performed on segments by SR-capable nodes. They are associated with operations performed on MPLS labels in MPLS networks. Segment Routing operations are: (a). PUSH – a segment is pushed on the top of segment stack (b) NEXT – an active segment is completed and it is removed from the stack (c). CONTINUE – active segment is not completed yet and it remains active. In Segment Routing network, it is enough to have an IGP protocol and once Segment Routing is configured, IGP will take labels and redistribute them within the domain.

In this paper, the switches update their forwarding tables according to the instructions taken from the controller. The switch informs the controller about the requested flow. The controller selects a path considering the requested bandwidth. After selecting the path, the controller sends flow information to the switches along the selected path using OpenFlow protocol. To determine the path, the controller needs to calculate the available bandwidth of the paths. For this purpose, the controller queries the switches periodically via sending OFPC_PORT_STATS messages which are defined in OpenFlow protocol to obtain information about available bandwidth on the links. When a traffic flow has to be routed along the shortest path to its destination, a segment list including only one label can be used (i.e., the SID of the destination node).



Node R3	
In Label	Action
104	pop,out 2
105	pop,out 3
106	pop,out 2

Node R4	
In Label	Action
103	pop,out 1
105	pop,out 1
106	pop,out 2

Node R6	
In Label	Action
103	pop,out 1
104	pop,out 1
105	pop,out 2

Node R5	
In Label	Action
103	pop,out 1
104	pop,out 1
106	pop,out 2

Figure 2. Example of Segment Routing

For the traffic from H1 to H2, if the controller selects the SR path P {R3, R4, R6}, a possible segment list used for P is SL= {106}. The packets are then forwarded along P without modifying the segment list to node R4 where the label 106 is popped and the packet is forwarded to node R6.

4. Quality of Service Provisioning

Quality of Service is an area of ongoing research and has been increased the interests in the research community. In today Internet, there are two main categories of QoS techniques: approaches proposed before SDN and SDN enabled approaches. When considering QoS approaches before SDN, the Internet Engineering Task Force (IETF) categories QoS into two major architecture: Integrated services (IntServ) and Differentiated Services (DiffServ). IntServ, per-flow design, specifies the elements to guarantee QoS and keep the network status information on every router. DiffServ, a class-based architecture, operates by classifying the traffic into classes with per-hop behavior. However, it cannot provide enough resources to guarantee QoS of different flows.

SDN enabled approaches to tackle all of the problems of the traditional network. SDN is a relatively new practice, and the cost of new technology is high. SDN controller can specify policies without the need to reconfigure low-level settings at each of the forwarding devices. The set of policies and the different flow classes are unrestricted. The rules can be defined per flow and the controller has the task to apply them properly to the different network elements. There are some QoS requirements of applications such as Bandwidth, hop-count, delay and jitter. In this system, the route selection of the flows is done by considering the paths from the controller to the switches. When a user needs a certain bandwidth, it sends a bandwidth request packet to the controller. Request packet contains information how much bandwidth it needs as a Packet-in message.

The SDN controller determines the segmented routed path in the network. When a new packet arrives at an OpenFlow switch, the switch will first check the packet header against all the preserved rules. If there is a match, then the switch will execute the matched rule action, otherwise, the network controller will be asked on how to deal with the incoming packet via receiving a *packet-in* request from the particular switch. Then, the controller will process the switch's request and respond by installing the proper rules through the *flow-mod* message.

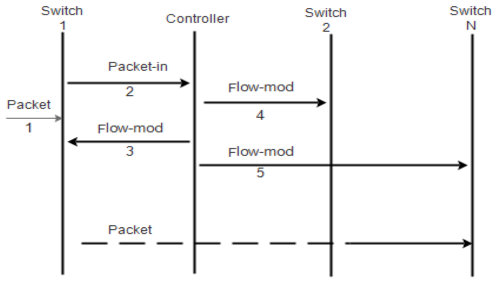


Figure 3. Interactions between Controller and OpenFlow Switch

Table 1. Notation Lists

Notation	Description
G(V,E)	The directed graph representation of the topology with vertex and edge
u_E	Utilization of link in the topology graph
C_E	The capacity of link
bw_u	The link usage bandwidth
bw_P	The available bandwidth of each path
$P_{S\text{->}D}$	The set of all available paths from S to D

The interactions between SDN controller and OpenFlow switch is as shown in Figure 3. This system uses this information to build up the network *G*(*V*, *E*) as shown in Figure 4, where the vertex *V* corresponds to the switches and the edge *E* corresponds to the links as shown in Table 1.

In this case, 'C' is the capacity of link, '*u*' is the link bandwidth utilization and bw_u is the bandwidth usage by monitoring the traffic flow of the link. For each link *E*, the available bandwidth resource of link E is $C_E - bw_u$. We define the available bandwidth of each path in Equation. (1):

$$bw_P = \min_{1 \le i \le n}(C_E - bw_u) \qquad (1)$$

Here, we have to get the path between source and destination in the network where the available bandwidth is the largest. This can be calculated through the following Equation. (2):

$$bw_a = \max_{P \in P_{S->D}} \min_{1 \le i \le n}(C_E - bw_u) \qquad (2)$$

Modified Dijkstra algorithm is used to get the path with largest available bandwidth. The cost of the path C_p is measured by the minimum bandwidth cost to obtain the best path according to Equation (3).

$$C_P = \sum_{i=1}^{n-1} C(bw_u, C_E) \qquad (3)$$

where the cost of the path C_p is the sum of the capacities of the link.

The routing algorithm is committed to find the best path for specific QoS requirements. A feasible bandwidth is the one that can provide sufficient resource to satisfy all QoS requirements of the flow. The algorithm is divided into two steps. The first step is to find the feasible bandwidth which can assure flow's QoS requirements, while the second step is to find a best-effort flow when feasible bandwidth doesn't exist. If feasible bandwidth exists, one of the paths will be selected to transmit the flow. To explain QoS routing Algorithm, consider a Mininet testbed for the network with four nodes shown in Figure 4. In the testbed, source node is S1 and destination node is S2. When the flows entered the network, for the case of higher bandwidth application flow S1-S3-S2, S1 would push a SR header with segment list *{*101,103,102*}*, and forward it to S2. Best-effort flow should be steered over the shortest path, which is S1 to S2. S1 would place the segment list *{*101,102*}* in the SR header, and forward to S2. Flow Classification is defined by Table 2.

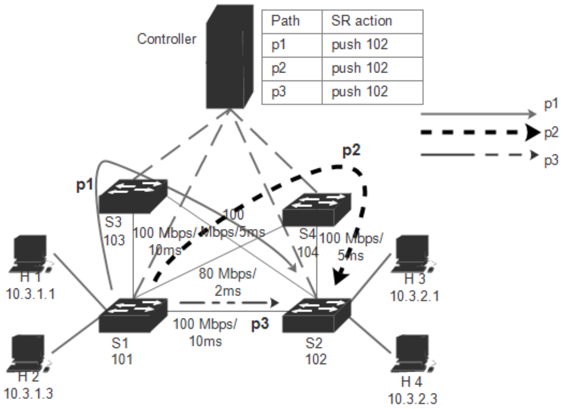


Figure 4. Test set up with Mininet

Table 2. Characteristics of Flows

Flow Type	Flow's QoS requirements	Flow
Minimum bandwidth flow (srp 1)	50 Mbps, < 10ms	H1-H3
Higher bandwidth flow (srp 2)	100 Mbps, < 20ms	H1-H3
Best-Effort flow (srp 3)	>100Mbps, -	H2-H4

```
//Create LeafSpine Topology//
Input: Switch, Host;
Initialize: S=0, H=0, N=4;
For all i=1 to N do
        $S_i$ =addSwitch ($S_i$);
        $H_i$ =addSwitch ($H_i$ , $IP_i$);
        $L_i$ = createLink ($S_i$);
Topo T = S,H,L;
Return T;
```

Figure 5. Algorithm of SR Topology

Flow type is obtained after flow classification as the basis to specify the QoS requirements. In this paper, it distinguishes the traffic to guarantee the bandwidth to three types. The first type is one to provide minimum bandwidth flow (srp 1) such as VoIP. The second type is set to use higher bandwidth flow (srp 2) such as video conferencing. The remaining type is set to best-effort flow (srp 3). The experiment is employed with Mininet testbed to evaluate QoS routing. We used the pseudo code to emulate the Mininet testbed of the network as shown in Figure 5. This section includes the setup and verification of SR test application, utilizing Mininet and the ONOS controller.

The pseudo code for our algorithm is summarized in Figure 6. Let G(V, E) be the network graph where the node set *V* corresponds to the switches and the edge set *E* corresponds to the links. Then srp defines segment routing (SR) path, bw_f defines the feasible bandwidth, bw_a represents available bandwidth and d_f is the feasible delay.

```
Initialize# Finding Available Bandwidth
Input:  Topology: G(V,E)
    Bandwidth Threshold $bw_{max}$; $bw_f$ ;
    Delay $d_f$ ;
    n= number of available paths,

Initialize: $bw_a$=0;
BEGIN
if ($bw_f \geq bw_{max}$)
    then $bw_a = bw_f$
        if ($bw_a \leq 50$ && $d_f < 10$ )
          then
                Add $bw_a$ to srp 1
                goto FINISH:
        else if ($50 < bw_a \leq 100$ && $d_f < 20$)
                Add $bw_a$ to srp 2
                    goto FINISH:
          else Add $bw_a$ to srp 3
end if
FINISH
for all n ∊ N do
   If $bw_n < bw_a$ || $bw_n < bw_{max}$ then $bw_n$ =0
endfor
```

Figure 6. Algorithm of QoS Routing

5. Experimental Results

We show the simulation results performed on different configurations to emphasize the capabilities and the performance of the QoS on a virtual machine (VM) with the configuration using Leaf-Spine architecture. This Leaf-Spine architecture is designed to provide very scalable throughput across thousands to hundreds of thousands of ports.

We created a Mininet testbed containing four nodes. The evaluation was performed on the Mininet testbed depicted in Figure 4. The script in this system is based on Python to generate flows. The Mininet testbed is comprised of 4 virtual switches interconnected with an SDN controller, 4 virtual hosts and virtual Ethernet links interconnecting the switches.

The SDN controller is an Open Network Operating System (ONOS) controller using OpenFlow configured to communicate with the data plane as shown in Table 3. The ONOS SPRING-OPEN project VM image has a 64-bit Ubuntu 16.04 installed as the guest OS. These are the minimum requirements to run the environment. The PC has Microsoft Windows 8.1 OS. The system was run with Core(TM) 1.6 GHz CPU and 4 GB of RAM. After collecting key performance parameters from both traditional network and SDN, this system model the data for a graphical representation.

In this system, the controller includes a routing policy based on a maximum bandwidth threshold 30Mbps between the switches. For this experiment, a Mininet testbed shown in Figure 4 is used. The requested services cannot provide if the request exceeds the threshold level of the link bandwidth. Srp 2 has a higher priority than srp 1 and we change the congestion level by injecting srp 3 into the network. The srp 1 and srp 2 having the higher level than srp 3 will have a higher priority queue to acquire sufficient bandwidth resource.

Hosts h1 and h2 send QoS traffic to h3 and h4 with the guaranteed rate of 30 Mbps. The actual rate sent by h1 is 100 Mbps and 50 Mbps respectively. Srp 3 rate between hosts is >100 Mbps. All flows are generated with iperf. The received throughput is observed from iperf's statistics.

Table 3. Parameters

Parameter	Values
Number of switches	4
Bandwidth threshold	30 Mbps
Delay sensitive threshold	20s, 10s
SDN controller	ONOS
Simulation tool	Mininet 2.2.1

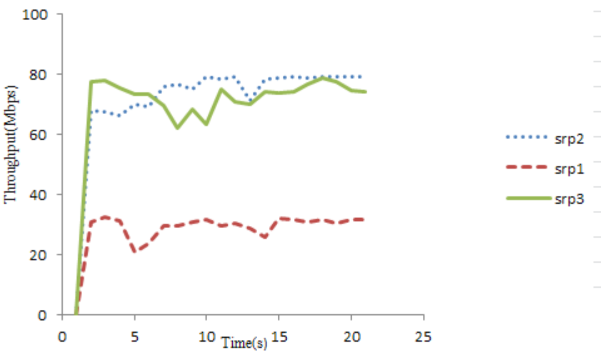


Figure 7. Throughput

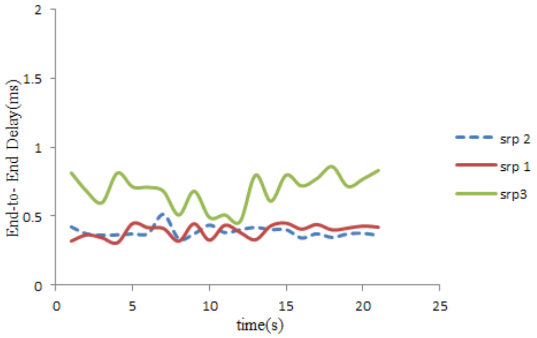


Figure 8. End-to-End Delay Variation

Figure 7 shows the time-varying throughput for srp 1, srp 2 and srp 3. We implement QoS control scheme at 15 s and detect the difference of throughput. Srp 1 uses the path S1-S4-S2 which is near 32 Mbps and Srp2 uses the path S1-S3-S2 which is 80 Mbps. Srp 1 and srp 2 are acceptable throughputs than srp 3 because srp 1 and srp 2 arrive at some peak rate at 20s. Congestion occurs when srp3 is at 10s and it is not available to guaranteed QoS. The routing algorithm has not acquired the proper route for srp 3 because of the constraint of delay.

The test result of delay variation is shown in Figure 8. The flows have suffered from huge variation at 10s. The srp 1 and srp 2 experienced at desired seconds. Srp 3 increases significantly at 10s due to congestion. The delay variation of srp 1 and srp 2 has decreased in 20s than srp3.

6. Conclusion

QoS provisioning is an important approach for several new architectures. This paper has implemented a solution to provide available bandwidth for QoS provisioning across SDN/OpenFlow network. This paper proposes QoS provisioning based segment routing over SDN environment to guarantee bandwidth efficiently for each application flow. In order to provide the QoS of higher bandwidth flow, segment routing is used to satisfy the requirement of service. The available bandwidth via experiments is implemented over SDN using Mininet testbed and ONOS controller. Based on the simulation results, the performance of our proposed QoS provisioning can improve the QoS requirements since the network throughput is stable compared with the throughput results in Best-Effort flows. In our future work, we plan to conduct extensive experiments with more complex network topologies to have more traffic by using OpenFlow Protocol.

7. References

[1] L.Chienhung, W.Kuochen and D.Guocin, "A QoS-aware routing in SDN hybrid networks". FNC 2017.

[2] S.Tomovic, I.Radusinovic and N.Prasad, "Performance comparison of QoS routing algorithms applicable to large-scale SDN networks". In EUROCON 2015-International Conference on Computer as a Tool (EUROCON), September 2015, pp. 1-6.

[3] H.Cho, J.Park, J.M.Gil, Y.S.Jeong, and J.H.Park, "An Optimal Path Computation Architecture for the Cloud-Network on Software-Defined Networking". Sustainability, 7(5), May 2015, pp. 5413-5430.

[4] U.Pongsakorn, I.Kohei, U.Putchong, D.Susumu and A.Hirotake, " Designing of SDN-Assisted Bandwidth and Latency Aware Route Allocation". (HPC), July 2014, pp. 1-7.

[5] T.T.Nguyen and D.S.Kim, "Accumulative-load aware routing in software-defined networks". In Industrial Informatics (INDIN), IEEE 13th International Conference, July 2015, pp. 516-520.

[6] P.Megyesi, A.Botta, G.Aceto, A.Pescapè and S.Molnár, "Available bandwidth measurement in software defined networks". In Proceedings of the 31st Annual ACM Symposium on Applied Computing, April 2016, pp. 651-657.

[7] P.Megyesi, A.Botta, G.Aceto, A.Pescapé and S.Molnár, "Challenges and solution for measuring avaiclable bandwidth in software defined networks". Computer Communications, 99, February 2017, pp. 48-61.

[8] C.Cetinkaya, E,Karayer, M.Sayit and C.Hellge, "SDN for segment based flow routing of DASH". In Consumer Electronics–Berlin (ICCE-Berlin), 2014 IEEE Fourth International Conference, September 2014, pp. 74-77.

[9] D.J.Hamad, K.G.Yalda and I.T.Okumus, "Getting traffic statistics from network devices in an SDN environment using OpenFlow". ITaS, 2015 Sep, pp. 951-956.

[10] R.Kumar, M.Hasan, S.Padhy, K.Evchenko and R.B,Bobba, "End-to-End Network Delay Guarantees for Real-Time Systems using SDN", 2017.

[11] S.Tomovic, I.Radusinovic and N.Prasad, "SDN control framework for QoS provisioning", 22nd Telecommunications forum TELFOR 2014, Serbia, Belgrade, November 2014.

[12] C. Filsfils, N.K. Nainar, C. Pignataro, J.C. Cardona, and P. Francois, "The segment routing architecture", In Global Communications Conference (GLOBECOM), December 2015, pp. 1-6.

Workshop Proceedings

The 2nd International Conference on Advanced Information Technologies – Workshop (**ICAIT 2018 Workshop**)

2nd November, 2018

Yangon, Myanmar

Organized by

University of Information Technology
Ministry of Education, Myanmar

Workshop Proceedings
The 2nd International Conference on Advanced Information Technologies – Workshop (ICAIT 2018 Workshop)

November, 2018

Contents

2nd November, 2018 (Friday)

Joint Workshop

Data Science and Data Mining

Distributed Computing

Image Processing

Internet of Things

Natural Language Processing

Software Engineering

Joint Workshop

Digital Signal Processing Laboratory at Saitama University

Biography

Tetsuya Shimamura is a Professor of Graduate School of Science and Engineering at Saitama University in Japan. He was Dean of Information Technology Center at Saitama University in 2014 and 2015. In 1995 and 1996, he joined Loughborough University, UK, and The Queen's University of Belfast, UK, respectively, as a visiting Professor. His research interests are in digital signal processing and its applications to speech, audio, image and communication systems. He has published over 100 refereed journal articles and 240 international conference proceedings papers. He is an author or co-author of eight books, and a member of the organizing committee of several international conferences. He has received IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Gold Paper Award, in 2012, WSEAS International Conference on Multimedia Systems and Signal Processing, Best Paper Award, in 2013, and IEEE IFOST, Best Paper Award, in 2014. Also, he is a recipient of Journal of Signal Processing, Best Paper Award, in 2013, 2015, and 2016.

Digital Signal Processing Laboratory at Saitama University

As one laboratory at Saitama University, activities of our laboratory are introduced. Our research area is digital signal processing and its applications to multimedia systems, which include wireless communication technologies. Students are separated into three research groups, which are speech group, image group and communication group. Research topics in each group are shown below.

Speech Group:
Spectrum estimation, pitch detection, speech enhancement, bone conduction, beamforming, speech or speaker recognition, watermarking etc.

Image Group:
denoising, quality assessment, restoration etc.

Communication Group:
spectrum sensing, channel estimation or equalization, modulation detection, interference suppression, PAPR reduction for OFDM etc.

In the above, some topics are picked up and explained in detail. Currently, many students are interested in speech and image processing systems. Approaches toward noise removal will be demonstrated.

In addition to research explanation, education systems at Saitama University are explained, where how to enter our graduate school is also shown. In many cases, international students are supported by a variety of scholarships to engage in studying in Japan. Related with this, practical states for our lab members are unveiled.

Furthermore, examples of research collaboration are shown, which include academics as well as companies.

Measuring and Presenting Material Appearance of Real Objects

Biography

Takashi Komuro received the B.E., M.E., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo in 1996, 1998, and 2001, respectively. At present he is an Associate Professor of mathematics, electronics and informatics at Saitama University. His current research interests include image sensing, computer vision, user interfaces, and augmented reality.

Measuring and Presenting Material Appearance of Real Objects

In online shopping, presentation of material appearance of products is important for enhancing consumers' willingness to purchase the products. Therefore, the system that reproduces realistic material appearance of real objects is required. For that purpose, we propose a method that measures shape and reflectance of real objects using simple apparatus. The system captures depth and color images of a rotating object on the turntable using an RGB-D camera. The shape of the object is reconstructed by integrating the depth images of the object captured from different viewpoints. The reflectance of the object is obtained by estimating the parameters of a reflectance model from the reconstructed shape and color images.

We also constructed a mobile AR system for presenting material appearance of objects. This system allows the user to manipulate a virtual object using his or her own hand with six degrees of freedom by mapping the hand pose to the object's translation and rotation. We measured the shape and reflectance of some real objects and presented the material appearance of the objects using the mobile AR system. It was confirmed that users were able to obtain the perception of materials from changes in gloss and burnish of the objects by rotating the objects using their own hand.

Internet of Things for the Masses

Biography

Dr. Yan Lin Aung received his Bachelor of Engineering (Computer Engineering) and Ph.D. in Computer Engineering from Nanyang Technological University, Singapore. Currently, he is a post-doctoral research fellow at iTrust, Centre of Research in Cyber Security, of Singapore University of Technology and Design (SUTD). His research interests include Internet of Things, cyber-physical systems and embedded systems security, runtime anomaly detection, embedded systems design and development, reconfigurable and custom computing.

Internet of Things for the Masses

Internet of Things (IoT) refers to systems of physical and virtual entities that are connected with one another, allowing interaction anytime and anywhere. The connectivity enables designers to make things 'smart' so it can track, log, display, monitor and adjust accordingly. It is anticipated that the emergence of IoT will transform nearly every global industry – from transportation, healthcare, agriculture, manufacturing, smart city, and many others. Bain & Company predicts that the combined markets of IoT will grow to about $520 billion in 2021. The workshop on "*Internet of Things for the Masses"* will provide a brief introduction to Internet of Things covering the architecture and key components of the technology. Realization of three applications – smart air quality monitoring, flood monitoring and automatic vehicle license plate recognition system, which will greatly benefit the masses, using the IoT technology will then be discussed. This workshop will also shed light on establishing a machine learning model based on the time series data from the sensors to make predictions of parameters. The presentation will be complemented with demonstrations.

Data Science and Data Mining

An Interrelation-based Approach to Aspect Extraction from Customer Reviews

Aye Aye Mar, Nyein Thwet Thwet Aung, Su Su Htay
Faculty of Information Science, University of Information Technology, Myanmar
{ayeayemar,nyeinthwet,suhtay}@uit.edu.mm

Abstract

Aspect extraction plays a key role in aspect-based sentiment analysis. Without knowing them, the target of sentiments expressed by customers cannot be known. If the target of the sentiments is not known, the sentiments expressed in a review are of limited use. Aspects and sentiment words are related to one another. Aspect extraction affects the performance of sentiment word extraction and also sentiment analysis. This study concentrates on aspects extraction from customer product reviews. This paper proposes an interrelation-based approach which considers the strength of interrelations between aspects and sentiment words to solve the problem of aspect extraction. The proposed approach takes into account the frequency of aspects, the weight of aspects and the weight of sentiment words associated with the aspects. Due to considering the interrelations between the aspects and sentiment words, the proposed method is expected to extract the relevant product aspects effectively and solve the major bottleneck of domain dependency.

Keywords- Aspects Extraction, Aspect-level, Feature Extraction, Sentiment Analysis, Opinion Mining, Customer Review, Text Mining

1. Introduction

Sentiment analysis of customer reviews has become a prominent research area during the last few years. Due to the rapid expansion of e-commerce, the web has become a source of grouping consumer opinions via customer reviews about the products. A collection of consumer opinions can be found in many product review websites such as amazon.com, epinions.com, cnet.com.

Sentiment analysis from customer reviews is vital for both customers and manufacturers to make the right decision. Customers inquire about the product they are interested via the reviews of other customers to decide whether he or she should buy that product or not. Alternatively, manufacturers improve the product quality and marketing campaigns based on the feedbacks of the consumers.

Sentiment analysis has been studied in three different levels of granularity: document level, sentence level and aspect level also called feature level [1]. Although sentiment analysis at document level and sentence level is useful in many applications, they cannot provide the necessary detail about the sentiment of the customers.

Among those three levels, aspect level is the most fine-grained level which extracts not only the opinions/sentiments but also the aspects/features/opinion targets. Unlike the other two levels, aspect-level sentiment analysis can decide what customers like and dislike. In the field of sentiment analysis, aspects are topics on which opinions are described. Other similar names for aspects are features, product features and opinion targets [2] [8].

Aspect-level sentiment analysis involves two tasks: aspect extraction and sentiment extraction. Aspect extraction is the most fundamental and important task of sentiment analysis because the sentiments expressed on those aspects are detected based on finding the aspects. So, this paper concentrates on the problem of aspect extraction from customer product reviews.

There are two types of aspects in aspect-level sentiment analysis: explicit aspects and implicit aspects [2]. Explicit aspects explicitly describe the name of targets in opinioned sentences e.g., "The weight of the phone is heavy". In that sentence, the weight is the aspect/opinion target and it is directly mentioned in the review. In contract, implicit aspects are not directly described in the opinion sentences and the aspects are expressed indirectly by using the implicit aspect indicator (sentiment word) e.g., "The phone is extremely light and disappears in your pocket". In that sentence, the aspect is not directly mentioned but it means weight.

Although both explicit and implicit aspects are important for sentiment analysis, explicit aspects are commonly occurred more than implicit aspects [2]. Explicit aspects extraction has been done by many previous researches but there has been limited number of research on implicit aspect extraction task.

This paper aims to solve the problem of explicit aspect extraction and proposes a method which considers the strength of relations between aspects and sentiment words. The proposed method take into account the frequency of aspects, the weight of aspects and the weight of the sentiment words associated with the aspects.

The rest of the paper is organized as follows: Section 2 summarizes the related work. The proposed system is presented in Section 3. In Section 4, the experimental evaluation matrixes are described. Section 5 concludes the paper and describes the future work of the proposed method.

2. Related Work

There have been many previous researches which have made aspect extraction for sentiment analysis using the supervised method, semi-supervised method and unsupervised methods [9]. The first and foremost attempt to aspect extraction was made by [2]. They introduced how implicit aspects differ from the explicit aspects and deal with explicit extraction problem. They used association rule mining based on the Apriori algorithm to extract frequent noun and noun phrases. They assumed that frequent noun and noun phrases as the explicit product aspects. To remove incorrect frequent aspects, they make feature pruning. Their approach considered only the frequency of noun and noun phrases so less frequent ones cannot be extracted. Our proposed method considers not only frequency of aspect candidates but also their weight and the weight of the respective sentiment words. Therefore, our approach is expected to be able to perform better than their work.

In [3], Somprasertsri proposed a supervised model to extract aspect by combining lexical syntactic features with maximum entropy technique. In their work, they presented four different features for learning maximum entropy. Those features include Aspects and their POS tags, Rare words, Alphanumeric feature and Dependency from syntactic parse tree. Those features were extracted from an annotated corpus. Maximum entropy classifier was used to extract the aspects.

Semantic-based product aspect extraction (SPE) method was presented by Wei et.al [4]. Their work used a list of positive and negative adjective of General Inquirer to recognize sentiment words semantically and subsequently extract aspects. They applied association rule mining to detect candidate product aspects. The used the same pruning strategy of [2]. They discovered the infrequent aspects by using semantic-based refinement. Their approach relies heavily on frequency and semantic-based extraction to detect aspects. In our approach, we study the interrelation information between the aspects and sentiment words by considering their weight.

In [5]**,** Poria et al. presented a rule-based approach to extract both explicit and implicit aspects. Rules were identified to extract aspects from product reviews. They defined implicit aspect clues (IAC) associated with implicit aspects. For identifying the aspects, they defined several dependency rules and applied WordNet and SenticNet to identify the synomyms and semantics of each IAC respectively. Their approach is limited to identify all possible rules among aspects and opinions. In addition, the association rule mining approach leads to the methodology computationally expensive.

The main goal of the proposed approach is to construct a domain-independent aspect extraction system which can extract aspects based on the interrelations between aspects and sentiment words.

3. Proposed System

The proposed system is mainly composed of three components: preprocessing, computing aspect cores for each aspect candidate and selecting the aspects from the aspect candidates. First and foremost, review dataset is preprocessed and Noun and Noun phrases are extracted as aspect candidates. Next, the system computes the aspect score for each aspect candidate and finally selects the aspects by using the threshold.

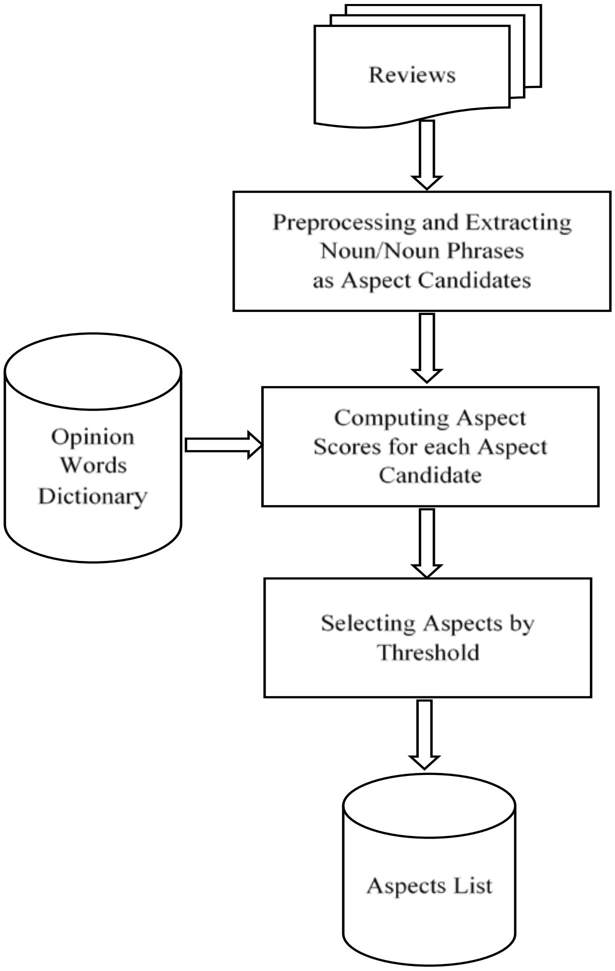

Reviews
Preprocessing and Extracting Noun/Noun Phrases as Aspect Candidates
Opinion Words Dictionary
Computing Aspect Scores for each Aspect Candidate
Selecting Aspects by Threshold
Aspects List

Figure 1. System flow of the proposed system

3.1. Preprocessing

Preprocessing is done before extracting the sentiment features. It includes two parts: POS tagging and stop words removal. POS tagging is done by applying the Standford POS tagger tool. POS tagging means labelling each word in a sentence with its appropriate part of speech such as noun, adjective, adverb etc.

Stop words such as verb to be, pronouns, prepositions and conjunctions do not give meaningful information for sentiment analysis. So, the stop words are removed to save the processing time.

3.2. Extracting the Aspects

Aspects are commonly found as noun or noun Phrases and they are occurred more often than the other noun and noun phrases. Accordingly to that observation, frequent noun and noun phrases are extracted as aspect candidates. To detect the relevant aspects from the aspect candidates, the proposed method computes the aspect scores by considering weight of relation between aspects and sentiment words.

The proposed method is based on the observation that there are interrelation between the aspects and the sentiment words. Interrelation information means the probability of co-occurrence of aspect and sentiment word in a sentence or in a review. The basic intuitions behind the proposed idea are as follows:

(i) An aspect is co-occurred with many sentiment words because different customers might describe their sentiments on the same aspect by using a variety of sentiment words (e.g. good camera, amazing camera, beautiful camera, lovely camera, bad camera, etc.,)

(ii) Alternatively, a sentiment word can be co-occurred with more than one aspect because customers often use some sentiment words to talk about their opinions about many aspects (e.g. good camera, good battery, good sound quality, low price, low processing time, etc.,)

Based on those assumptions, the proposed method computes the scores of aspects by considering frequency of aspects, the weight of aspects and the weight of the sentiment words associated with the aspects. The frequency of aspects is the number of occurrence of aspects in the reviews. The weight of an aspect means the number of distinct sentiment words that is co-occurred with that aspect in the reviews. Similarly, the weight of a sentiment word is the number of frequent noun or noun phrases (other aspect candidates) that is co-occurred with that sentiment words.

The Algorithm 1 describes about aspect extraction process from POS tagged reviews. A number of sentiment word dictionaries are available in recent researches. SentiWordNet has a wide application in the field of sentiment analysis and it has the largest number of features, and its structure is suitable for mathematical modeling [7]. Therefore, SentiWordNet is used to detect the sentiment words from the reviews [6].

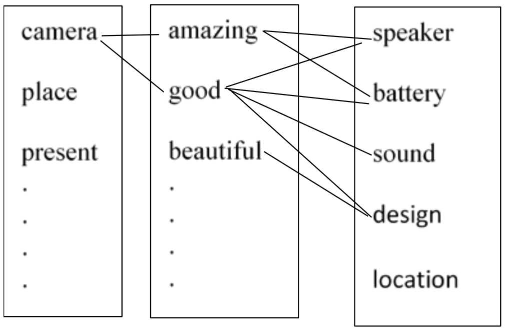


Figure 2. Relations between aspects and sentiment words.

In Figure 2, camera, place and present are aspect candidates. Amazing, good and beautiful are sentiment words which is contained in SentiWordNet dictionary and co-occurred with the aspect candidate camera. The weight of aspect camera is 2 because it is found together with two sentiment words: amazing and good. The system considers the weight of those two sentiment words so it detects other aspect candidates which are not aspect candidate camera. In that picture, the weight of sentiment word amazing is 2, the weight of sentiment word good is 4 and the weight of sentiment word beautiful is 1. To compute the aspect score of aspect camera, the system will consider the frequency of camera, the weight of camera and the weights of its associated sentiment words (amazing, good, beautiful). The more interrelation between aspects and sentiment words, the more possible for that aspect candidate to be aspect.

4. Evaluation Matrix

To evaluate the performance of the proposed method, four evaluation metrics: precision, recall, F-measure and accuracy are considered to evaluate the effectiveness of the system. These are calculated by using Eq. (1) - (4) respectively.

$$\Pr ecision = \frac{TP}{TP + FP} \quad (1)$$

$$\operatorname{Re} call = \frac{TP}{TP + FN} \quad (2)$$

$$F - measure = \frac{2 * \Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

where:
TP refers to the number of true positive reviews.
TN refers to the number of true negative reviews.
FP refers to the number of false positive reviews.
FN refers to the number of false negative reviews.

Algorithm 1. Aspects Extraction Algorithm

```
Input : POS tagged Review Document D,
        opinion-word-dictionary, threshold
Output : List of aspects

for each aspect candidate a (N/N phrase) in D do
    for each ow in opinion-word-dictionary do
        for each sentence in D do
            if ow is found together with a then
                $Weight_{a=} \; Weight_{a} + 1$
                if ow in not in ow-list-for-a then
                    add ow into ow-list-for-a
                end if
            end if
        end for
    end for
    for each opinion word ow in ow-list-for-a do
        for each sentence in D do
            if N/N phrase (not a) found near ow then
                $Weight_{ow} = Weight_{ow} + 1$
            end if
        end for
    end for
    Compute $f_a$ (the frequency of a )in D
    $a\text{-}score = log (f_a * Weight_a * \sum_{ow \in ow\text{-}list\text{-}for\text{-}a} Weight_{ow})$
    if (a-score > threshold ) then
        add a into $List_{aspect}$
    end if
end for
Return $List_{aspect}$ as the list of aspects
```

5. Conclusion

In this paper, we study the nature of the aspects and sentiment words and propose an aspect extraction method based on the interrelation between the aspects and sentiment words. Since the proposed method is based on relation of the words, the method will extract the relevant aspect and solves the major bottleneck of domain dependency.

In our future work, we will make evaluation of the proposed method firstly with the product review datasets and then with the datasets from another domains. We have plan to make comparative evaluation of the proposed method by using the domain-specific sentiment dictionary instead of SentiWordNet.

6. References

[1] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[2] Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).

[3] Somprasertsri, G., & Lalitrojwong, P. (2008, July). Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features. In Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on (pp. 250-255). IEEE.

[4] Wei, C. P., Chen, Y. M., Yang, C. S., & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. Information Systems and E-Business Management, 8(2), 149-167.

[5] Poria, S., Cambria, E., Ku, L. W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In Proceedings of the second workshop on natural language processing for social media (SocialNLP) (pp. 28-37).

[6] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

[7] Khan, F. H., Qamar, U., & Bashir, S. (2017). A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. Knowledge and information Systems, 51(3), 851-872.

[8] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[9] Rana, T. A., & Cheah, Y. N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. Artificial Intelligence Review, 46(4), 459-483.

Myanmar Stock Market Prediction Model using Multi-view Feature Selection

Ei Thwe Khaing
University of Information Technology,
Yangon, Myanmar
eithwekhaing@uit.edu.mm

Myint Myint Thein, Myint Myint Lwin
Faculty of Information Science
University of Information Technology,
Yangon, Myanmar
myintmyintthein@uit.edu.mm,
myintmyintlwin@uit.edu.mm

Abstract

Stock Market Prediction is an essential task for traders and investors in the financial markets. Early market prediction used news articles and social media data to predict stock price direction. Nowadays, features in global conditions, political situations, and economic factors are utilized for stock prediction. Features in these factors are formed into high dimensional features that accuse overfitting problem. For high dimensional features, multi-view predictions produce better results than single view predictions. This paper proposes stock market prediction model based on multi-view feature selection to provide stock price direction. Stock market prediction system is used as an application in the proposed system. The proposed model has two stages on stock data. In the first stage, features are extracted from each factor as single view by using principal component analysis method. In the second stage, canonical correlation analysis method is applied to integrate features that are extracted from first stage. This prediction model intends to predict the stock price direction for Yangon Stock Market in Myanmar.

Keywords- Stock market prediction, Multi-vew feature selection, Principal component analysis, Canonical correlation analysis

1. Introduction

Prediction of stock price direction may provide traders and investors good profit in economy and stock market prediction. Many research studies have utilized historical stock price data as features for prediction. Nowadays, the prediction of stock price direction has several factors such as global conditions, political situations and economic indicators in financial markets. In the literature, technical analysis and fundamental analysis are used to predict the future stock price direction. Technical analysis uses historical stock price to predict future price direction (eg. past weekly stock data information). Fundamental analysis tries to make predictions based on the structure of economy (eg. inflation rates, exchange rates, trading volume).

News articles and social media provide related information about stock market behaviors. Many researchers analyze stock market prediction on news articles and social media. Machine Learning, Data Mining or Artificial Neural Networks methods applies to remove unrelated information and to extract related information on stock prices. This paper proposes stock price prediction model from local news, social financial web pages and government announcements in Myanmar.

The high dimensional financial data are collected from local news articles, financial web pages and government announcements within a week and stock price direction from stock market. And then the proposed model predicts the direction of stock price next week. Text mining techniques are used to transform the news articles into feature vectors for trading next week. There are three class labels for stock price direction, +1, 0 and -1, which define increase, no change and decrease, were assigned to these feature vectors according to the stock price change of the Yangon Stock Market.

The reminder of the paper is organized as follows. Section 2 presents an overview of related work for stock market predictions. Section 3 proposes Myanmar stock market prediction system using multi-view feature selection. Section 4 concludes this paper.

2. Related Work

In recent years, there have been many research studies on the prediction of stock price direction. The stock market prediction is surveyed by using the financial news. Each new article is separated to classes as price direction (increase, decrease or no change) and then a model is trained with these articles. The stock market prediction is predicted by performing result on the models with the recently released news articles.

The systematic approach for the stock market prediction developed to predict short-term stock price on the non-linear, volatile and complex nature of the market. The neural networks, support vector regression and the boosting used as a base learner. The models using the features from these external sources along with the traditional stock market data improved the performance

The daily stock movements presented on three stocks (GARAN, THYAO and ISCTR) in Borsa Istanbul (BIST) using different types of technical indicators as features. The filter feature selection and gradient boosting machine applied on expanding feature space on BIST stock features with noisy or irrelevant features.

The trend of stock market was very complex and influenced various factors in the trend of stock market on Shanghai Stock Exchange Composite Index (SSECI). Feature selections methods, such as principle component analysis, genetic algorithms and sequential forward search, used to find out the most significant factors to the stock market and to remove the redundant and irrelevant factors.

The daily price direction was forecasted by using classification mining procedure on S&P 500 Index ETF. Principal Component Analysis (PCA), fuzzy robust principal component analysis (FRPCA) and kernel-based principal component analysis (KPCA) applied to clean and complete data in order to select the most influential and uncorrelated variables for classification. Artificial Neural Network (ANNs) as classifiers used with the transformed data sets to forecast the direction of future market returns.

The Internet of Multimedia of Things (IMMT) for stock analysis demonstrated at Chinese stock market Forecasting. The deep long short-term memory neural network method used in embedded layer with automatic encoder. The method improved the predicted effect of Shanghai A-share composite index to a certain factor with deficient historical data.

3. The Proposed Stock Market Prediction System

Stock market prediction is the act of trying to determine the future value of a stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The prediction model forecasts future price direction based on past historical financial data. This system contains two stages to construct a stock price prediction model. Stock data was collected from news and used to predict price direction. The successful stock market prediction uses several financial factors (global conditions, political situations, and economic factors) about price changes. Firstly, the relevant stock features are extracted from local news articles, financial web pages and official sites, as several factors. Secondly, the extracted subsets of features from each factor are combined and selected appropriate stock information. Finally, the information is used to construct a stock prediction model.

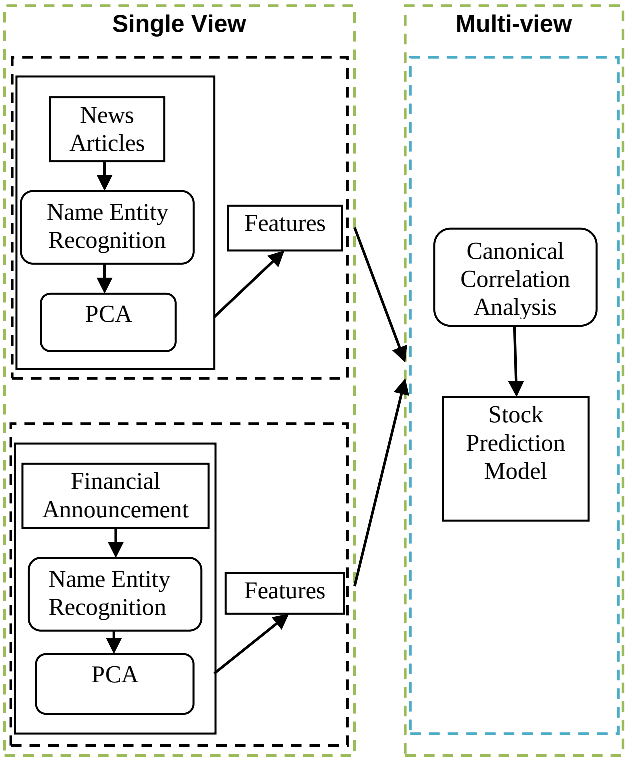


Figure 1. Myanmar Stock Prediction System

3.1. Single View

Many statistical techniques such as data mining, machine learning are used to extract relevant and remove redundant features/information in data analysis. Single view feature extraction leads to consider data on a specific dataset. There are three parts in this view. First, data are collected from local news and social media data for stock price prediction. Second, raw data tokenizes and remove unrelated word in data preprocessing. Third, potential features are extracted to predict the stock direction from a specific view/dataset.

3.1.1. Data Collection. Data are collected from news (Global New Light of Myanmar, The Myanmar Time), social media web sites/pages (Agriculture and Market Information Agency (AMIA), The Farmer Media, Ba Yint Naung and Ministry of Commerce).

Figure 2 is an example of local business articles from Global New Light of Myanmar.

> *The price of a ton of the same crop decreased to Ks430,000 on August 1, Ks426,000 on August 2 and Ks423,500 on August 3. Local consumption for this variety of crop is high and the prices of mung beans are traditionally higher than the pigeon peas (red grams) in the domestic market. The prices of pigeon peas were significantly higher than that of mung beans.*

Figure 2. A sample new article in AMIA

Figure 3 is an example of government announcement of Vice President U Henry Van Thio delivered the speech at the workshop on developing trade and export of Myanmar pulses, beans and sesame in Yangon on August 18, 2018.

> *Myanmar is an agriculture nation and rice, pulses and beans are the main agriculture products. Rice is the staple food of the country and is exported only after there is enough for local consumption. But only some pulses and beans are consumed locally and most are exported to foreign markets and up to fiscal year 2016-2017, it was the crops that earned the most foreign exchange.*

Figure 3. An announcement of Vice President

3.1.2. Data Preprocessing. Data preprocessing is the transformation of raw text into a collection of words. This process consists of tokenization, stop word removal and stemming. In the transformation, Bag of Words, Noun Phrases and Named Entities are used for text representations that are applied text to convert them. Name Entity Recognition is one of the useful information extraction techniques to identify and classify named entities in text. It extracts information from unstructured text data and categorizes it into groups such as persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Name Entity Recognition is applied in this part to tokenize words from news articles and financial data in social or government web sites. Next, words with low predictive capability are removed from these documents. And then stemming words (different words in same meaning) are transformed to a root form. Finally, a collection of words after preprocessing on data in Figure 2 is shown in Figure 4.

> *"crop on August 1", "decrease", "Ks430,000", "crop on August 2", " Ks426,000", "crop on August 3", " Ks423,500", " Local consumption", "high", "mung beans", "pigeon peas", "domestic market", " significant"*

Figure 4. A collection of words from text transformation

3.1.3. Single view Feature Extraction. Feature extraction transforms the original vector space into a new one with some special characteristics and the reduction is made in a new vector space. They are transformed (using a linear or non-linear transformation) to a reduced dimension space with the aim of replacing the original features by a smaller set of essential features.

Principal component analysis (PCA) is effective in data compression and feature extraction. PCA in text categorization use to get the low-dimension representation of document vectors. PCA is also useful when there are a large number of correlated dimensions that contain a lot of data redundancy. PCA can be employed to reduce this redundancy, which results in reduction of highly correlated data into small number of un-correlated principal components.

In principal components analysis, a multivariate data matrix **X** contains with *n* rows and *p* columns. The n numbers of stock features are contained. The *p* elements of each row are stock price direction such as increase, no change, and decrease. As a single view, a correlation matrix constructs by using features from the related stock features and price direction in describing Figure 4. Another matrix is constructed based on extracted stock price prediction information from government announcements. Table 1 shows a correlation matrix with the extracted stock features and price direction from government announcement. Another correlation matrix that relates the extracted features and price direction from news articles is described in Table 2.

Table 1. A correlation matrix for government announcement

	Increase	No Change	Decrease
rice	0	1	0
bean	1	0	0

Table 2. A correlation matrix for news articles

	Increase	No Change	Decrease
Crop on August 1	0	0	1
Crop on August 2	0	0	1
Crop on August 3	0	0	1
Crop	1	0	0
Mung bean	1	0	0
Pigeon pea	0	1	0

Principal component analysis is a nature of dimension reduction. It reduces a large set of features to a small set when correlated features contain in matrix. Therefore, the matrix in Table 2 reduces by combining the correlated features. The new transformed matrix result is mentioned in Table 3.

Table 3. A correlation matrix by dimension reduction

	Increase	No Change	Decrease
Crop	0	0	1
Mung bean	1	0	0
Pigeon pea	0	1	0
Mung bean	0	1	0

3.2. Multi-view

Multi-view uses the multiple views of data to provide predictable correlated features for the uncorrelated features for each other. The all of subsets of features in the several views/factors are integrated in this stage. Therefore, this stage selects the optimal subset of features from feature sets of each view. Feature selection approach chooses a subset of features from the datasets and aims to minimize feature redundancy and maximize the feature relevance to the target class label. The extracted features from each single view are integrated as the maximum relevant and minimum redundant features from multiple views.

3.2.1. Multi-view Feature Selection. Canonical Correlation Analysis (CCA) is widely used for feature selection on multiple data sources. CCA is a statistical method for finding correlation relationships between two sets of features (two views). CCA is a way of inferring information from cross-covariance matrices. If we have two vectors and there are correlations among the variables, then CCA will find linear combinations of these vectors which have maximum correlation with each other.

The combinations of two vectors, one of stock direction vector from government announcements in Table 1 and another vector from local news and social media data in Table 3, are highly correlated. The correlation between news articles and announcements is displayed in Table 4. The correlation between crop and price direction is decreased. The price direction of mung bean is increased. The correlation of pigeon pea and price direction is no changed.

Table 4. Correlation matrix between two views

	Increase	No Change	Decrease
Crop	0	0	1
Mung bean	1	0	0
Pigeon pea	0	1	0

4. Conclusion

Stock market prediction is a model to predict stock price movements and to improve stock investments based on historical stock data information. In this paper, stock market prediction model is proposed for price direction. It provides Yangon stock market prediction that effects on financial data. The stock price information in previous week's news articles, web pages and announcements is extracted by applying the principal component analysis as a single view. The canonical correlation component uses to select the optimal subsets of features from multiple views. Finally, a stock market prediction model constructs to predict next week stock price direction over the multi-view feature selection.

5. References

[1] E.M. Yasser, "CCA based multi-view feature selection for multi-omics data integration", *bioRxiv*, 2018, p.243733.

[2] H. Gunduz, Z. Cataltepe, and Y. Yaslan, "Stock daily return prediction using expanded features and feature selection", *Turkish Journal of Electrical Engineering & Computer Sciences*, *25*(6), 2017, pp. 4829-4840.

[3] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification", *Engineering Applications of Artificial Intelligence*, *52*, 2016, pp.26-39.

[4] S. Chormunge, and S. Jena, "Correlation based feature selection with clustering for high dimensional data", *Journal of Electrical Systems and Information Technology*, 2018.

[5] X. Pang, Y. Zhou, P. Wang, W. Lin and V. Chang, "An innovative neural network approach for stock market prediction", *The Journal of Supercomputing*, 2018, pp.1-21.

[6] X. Zhong, and D. Enke, "Forecasting daily stock market return using dimensionality reduction", *Expert Systems with Applications*, *67*, 2017, pp.126-139.

[7] Y. He, K. Fataliyev, and L. Wang, "Feature selection for stock market analysis", In *International Conference on Neural Information Processing*, Springer, Berlin, Heidelberg, November 2013, pp. 737-744.

[8] Y.M. Xu, C.D. Wang, and J.H. Lai, "Weighted multi-view clustering with feature selection", *Pattern Recognition*, *53*, 2016, pp.25-35.

Proposed CS-index and Querying Approach for SPARQL Queries

Khin Myat Kyu
University of Information Technology, Yangon
khinmyatkyu@uit.edu.mm

Kay Thi Yar
University of Information Technology, Yangon
kaythiyar@uit.edu.mm

Aung Nway Oo
University of Information Technology, Yangon
aungnwayoo@uit.edu.mm

Abstract

The interlinking nature of web-scale RDF data pose a challenge to storage and retrieving of these data efficiently. Even though different storage and query processing techniques have been proposed, query processing on relation-based RDF stores requires many join operations when the input query is complex (with respect to number of triple patterns). One solution to this problem is to reduce the number of joins by indexing. Indexing is an effective technique to reduce data searching space and retrieve data as fast as possible. In this paper, we propose new indexing scheme for chain and star queries (CS-index) and querying approach. The proposed approach could support both chain and star query. Our approach employs graph pattern based technology: the RDF data graph is firstly decomposed into chain and star shaped subgraphs based on the structural information of each vertex. These subgraphs are stored as index, called CS-index. When a SPARQL query is given, it is decomposed into query subgraphs based on common join variable among all triple patterns. And the query results are finally obtained by matching these query subgraphs against with CS-index. The proposed approach tends to minimize the query execution time by reducing the number of join operations as well as reduce memory usage for storing data.

.

Keywords- RDF, SPARQL query, graph-based index

1. Introduction

Resource Description Framework (RDF) is a flexible, schema-free and graph-structured data model for describing resources on the Web. Resources may be person, places, organizations, or anything on the Web. These RDF data can be accessed by SPARQL is a declarative query language recommended by W3C. As the highly interconnected nature of Web data, many RDF data management systems have been proposed with different techniques [1], i.e., relation-based RDF store, the clustered property table, vertical partitioning, multiple indexing. Relation-based RDF stores such as Jena-SDB, Sesame, RDF-3X, manage the data in relational tables and process SPARQL queries using relational operators, such as scan and join operators. The main problem of relation-based RDF stores is that they need too many join operations for processing SPARQL queries, especially for complex queries. Many approaches have been proposed to solve this issue by emphasizing on: (i) reducing number of join, (ii) reducing inputs of join operators, and (iii) optimizing join order [3]. However, the graph-structured nature of the RDF and the graph pattern matching nature of SPARQL queries still have significant challenges for efficient processing of complex SPARQL queries over the interlinking RDF data. A question arises to ask how to find efficiently all matches of a query graph in a large database graph, i.e., reducing time of query processing as much as possible.

In this paper, we propose new indexing and querying approach for processing chain and star query. Blank nodes are not considered as they represent a resource without specifying its URI. Chain query is a SPARQL query consist of subject-object joins where the triple patterns are consecutively connected like a chain. A SPARQL query consists of subject-subject join where each join variable is the subject of all the triple patterns involved in the query is called star query. Since the goal is to minimize query processing time and memory space for data storage, the proposed indexing structure and query processing algorithm uses (i) key-value pair based storage style, and (ii) dictionary encoding to minimize storage space more efficiently, and (iii) an indexing algorithm based on the combinations of vertices and predicates. The proposed method could minimize query execution time, and requires little memory usage as it stores all predicates of one vertex with one key-value pair.

The remainder of this paper is organized as follows: Section 2 describes literature review and explanation of RDF data and SPARQL query is given in Section 3. The proposed method is presented in Section 4. Section 5 explains the query evaluation with proposed method. Finally, Section 6 concludes the whole paper and discusses future perspectives.

2. Related Work

A common approach for storing and indexing RDF data used by many systems i.e., RDF-3X [4], Hexastore [5] is to store all triples (S,P,O) in a single three-column table. For efficient data access, one-dimensional indexes

(B+ trees) are used for each of the six SPO permutations (i.e., SP, SO, PS, PO, OS, OP), known as sextuple indexing technique. However, this querying efficiency comes at the cost of excessive storage requirements and maintenance overhead since the complete data set is stored replicated six times. It degrades the efficiency of query processing as it requires expensive self-joins when SPARQL queries consist multiple triple patterns [1].

X. Wang et. al [6] proposed a RDF storage and indexing scheme, called CHex. CHex uses sextuple indexing and *binary association table* (BAT) for a column-oriented database system. It not only provides efficient single triple pattern lookups, but also allows fast merge-joins for any pair of two triple patterns. But the additional processing becomes substantial as the queries become complex. And it incurs space overhead in data storing.

X. Lyu et. al [7] proposed the efficient subgraph matching method for star queries. The method decompose both data graphs and query graphs into sets of star graphs, and encode each star subgraph into a fingerprint. Fingerprints were used to effectively reduce the data searching space. But the method takes too much time in fingerprints encoding, and can handle only star queries.

In [8], RP-filter was proposed for reducing the redundant intermediate results of join operations. RP-filter uses a path-based index which indexes the incoming path information of RDF graph. However, it has limitation that it could not exploit the graph structural information of RDF data. The additional processing becomes substantial as the queries become complex. In order to overcome this limitation, RG-index was proposed in [9]. The RG-index indexes the graph patterns by using adapted gSpan algorithm - is a frequent subgraph mining algorithm was originally proposed for graph transaction data set eg. chemical compounds. But the method takes too much time for mining discriminative and frequent graph patterns from RDF data. To reduce time overhead, we propose CS-index which extracts chain and star shaped graph patterns by counting the incoming and outgoing degrees of vertices while parsing and dictionary encoding. We assume that extraction of chain and star shaped patterns need the time than RG-index because our proposed method does not use subgraph mining technique.

3. Preliminary Concepts

In this section, we present the concept of RDF data and SPARQL query. Assume that there are three pairwise disjoint sets: a set of uniform resource identifiers (URIs) U, a set of literals L, and a set of variables VAR.

3.1. RDF Data

A RDF data set is a collection of statements in the form of subject (s), predicate (p), and object (o). A statement t ϵ U × U × (U υ L) (without variables) is called a triple. Table 1 presents an example of RDF data set, LUBM - is a standard data set which was developed to evaluate the performance of Semantic Web repositories. For simplicity, we use prefix for each URI as large amount of triples share the same URI.

Prefix:
rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
rdfs = "http://www.w3.org/2000/01/rdf-schema#"
owl="http://www.w3.org/2002/07/owl#"
ub = "http://swat.cse.lehigh.edu/onto/univ-bench.owl#

Table 1. Example RDF data set

Subject	Predicate	Object
ub:Peter	rdf:graduatedFrom	ub:MIT
ub:Peter	rdf:advisor	ub:Lisa
ub:Lisa	rdfs:type	Professor
ub:Peter	rdf:takes	ub:DataMining
ub:Peter	rdf:takes	ub:Algorithms
ub:Peter	rdfs:type	GraduateStudent
ub:MIT	rdfs:type	University
ub:James	rdfs:type	Professor
ub:James	rdf:worksFor	owl:CS
owl:CS	rdf:subOrganizationOf	owl:InfoLab
owl:CS	rdfs:type	Department
ub:James	rdf:memberOf	owl:InfoLab
ub:James	rdf:teaches	ub:Algorithms
ub:Algorithms	rdfs:type	owl:Course
ub:Tim	rdf:takes	ub:Algorithms
ub:Tim	rdf:takes	ub:DataMining
ub:DataMining	rdfs:type	owl:Course
ub:Tim	rdf:advisor	ub:Fred
ub:Fred	rdf:teaches	ub:DataMining
ub:Fred	rdf:memberOf	owl:InfoLab
ub:Fred	rdfs:type	AssociateProfessor
ub:Fred	rdf:worksFor	owl:CS

These RDF data can be represented with directed, labeled graph. Figure 1 shows the RDF graph for example RDF data set in Table. 1. In this paper, blank nodes are omitted as they represent a resource without specifying its URI.

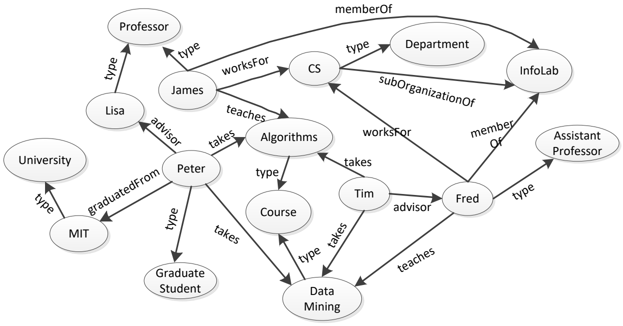


Figure 1. RDF graph

3.2. SPARQL Query

A SPARQL query consists of one or more triple patterns (tps). A statement tp ϵ (U υ VAR) × U × (U υ L υ VAR) (triple with variables) is called a triple pattern. Variable symbols start with ‘‘?’’ to distinguish them from URIs and literals. SPARQL queries can be classified based on the shape of the query graph. In this paper, chain and star query are considered. Chain queries include subject-object join (the join is between a tp's subject and another tp’s object). A star query includes subject-subject join, i.e. join variable is at the subject’s position of all the tps.

Figure 2. shows example of SPARQL query. It retrieves the university’s name where Peter graduated. It consists in the chain query type.

```
SELECT ?uni
WHERE { ub:Peter rdf:graduatedFrom ?uni.
        ?uni rdfs:type “University” . }
```

Figure 2. Example SPARQL query

The query graph of the example SPARQL query in Figure 2 is shown in Figure 3.

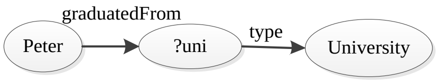


Figure 3. Example SPARQL query graph

4. Proposed Method

In our proposed method, there are two main phases: (i) index construction and (ii) query processing. The proposed indexing and querying algorithm are designed to process both chain and star query. Pseudo code for the proposed algorithms are described in Figure 5 and Figure 6, respectively.

4.1. Index Construction

There are some tasks in CS-index construction phase: parsing RDF triples, constructing dictionaries, and extracting graph patterns based on the incoming and outgoing degree of each vertex. All these tasks are carried out in parallel. As first task, each RDF triple (v_i,e_i,v_j) is parsed into three parts: subject, predicate, and object. Values of the subject/predicate/object are URIs or literals. Thus, we store integer values instead of these URIs and literals because they are complex and long string values.

Two dictionaries are needed for mapping the RDF triples with integer values. The first one is for subjects and objects, and the other one is for predicate, called subject/object dictionary and predicate dictionary, respectively. Key-value (id, value) mappings are used to construct the dictionaries. The notation for ‘id’ is defined as V_{id} and E_{id} where V_{id} is the integer value for subjects and objects, E_{id} is the integer value for predicates.

V_{id}	URI/Literal
1	ub:Peter
2	ub:MIT
3	ub:Lisa
4	Professor
5	ub:DataMining
6	ub:Algorithms
7	GraduateStudent
8	University
9	ub:James
10	owl:CS
11	owl:InfoLab
12	Department
13	owl:Course
14	ub:Tim
15	ub:Fred
16	AssociateProfessor

(a)

E_{id}	URI
1	rdf:graduatedFrom
2	rdf:advisor
3	rdfs:type
4	rdf:takes
5	rdf:worksFor
6	rdf:subOrganizationOf
7	rdf:memberOf
8	rdf:teaches

(b)

Figure 4. (a) Subject/object dictionary, (b) Predicate dictionary

While constructing the dictionaries, in-degree and out-degree are computed for each subject and object. Degree computation is not need to consider for predicate. If the parsed one (v_i) is subject, we increase the out-degree and add the pair (e_i,v_j) into outgoing-edges of v_i. If it is object, we increase the in-degree and add the pair (v_j,e_i) into incoming-edges of v_i. If it is predicate, next triples is read to parse.

Algorithm 1: CS-index construction algorithm

```
Input: RDF data set D
Output: CS-index, subject/object dictionary,
        predicate dictionary
```

```
begin
for each triple t in D
  parse and encode each URI/literal
  for each encoded URI/literal $v_i$
    if encoded URI/literal $v_i$ is subject
      if out-degree of $v_i$ is zero
        out-degree of $v_i$ ++
        outgoing-edges of $v_i$ = {($e_i$, $v_j$)}
      end if
      else
        out-degree of $v_i$ ++
        merge ($e_i$,$v_j$) to existing outgoing-edges of $v_i$
      end else
    end if
    else if encoded URI/literal $v_i$ is object
      if in-degree of $v_i$ is zero
        in-degree of $v_i$ ++
        incoming-edges of $v_i$ = {($v_j$, $e_i$)}
      end if
      else
        in-degree of $v_i$ ++
        merge ($v_j$, $e_i$) to existing incoming-edges of $v_i$
      end else
    end else if
    else break;
  end for
end for
for each encoded subject/object vertex $v_i$
  store outgoing-edges and incoming-edges,
  and $v_i$
end for
end
```

Figure 5. Pseudo code for index construction

After all RDF triples have been processed completely, we store incoming-edges and outgoing-edges as compound key and v_i as value in CS-index. And then, CS-index is sorted in ascending order based on the in-degree and out-degree pair of each v_i. CS-index of example RDF data in Table 1 is shown in Table 2. First column is used as line number for explanation in query evaluation section.

Table 2. CS-index architecture

	outgoing-edges (e_i,v_j)	incoming-edges (v_j, e_i)	V_{id}
#1	-	{(1,3)}	7
#2	-	{(2,3)}	8
#3	-	{(10,3)}	12
#4	-	{(15,3)}	16
#5	-	{(3,3),(9,3)}	4
#6	-	{(5,3),(6,3)}	13
#7	-	{(10,6),(9,7),(15,7)}	11
#8	{(3,8)}	{(1,1)}	2
#9	{(3,4)}	{(1,2)}	3
#10	{(3,13)}	{(1,4),(14,4),(15,8)}	5
#11	{(3,13)}	{(1,4),(9,8),(14,4)}	6
#12	{(6,11),(3,12)}	{(9,5),(15,5)}	10
#13	{(4,5),(4,6),(2,15)}	-	14
#14	{(3,4),(5,10),(7,11),(8,6)}	-	9
#15	{(8,5),(7,11),(3,16),(5,10)}	{(14,2)}	15
#16	{(1,2),(2,3),(4,5),(4,6),(3,7)}	-	1

4.2. Query Processing

When there is a query arrives, the query processor finds common join variable which include as a variable in more than one triple patterns. And the triple patterns are grouped based on the common join variable. And each subject/predicate/object values are encoded using two dictionaries constructed in index construction stage. Then, in-degree, out-degree, incoming-edges, and outgoing edges are computed for each common join variable.

```
Algorithm 2: Query processing algorithm
Input: SPARQL query Q, CS-index
Output: result of the query $result_Q$
begin
  find common join variable $var_i$
  decompose triple patterns tps based on $var_i$
  compute in-degree, out-degree, incoming-edges,
      outgoing-edges of $var_i$
  $result_Q$ = match($in\text{-}degree_{vari}$, $out\text{-}degree_{vari}$,
      $incoming\text{-}edges_{vari}$, $outgoing\text{-}edges_{vari}$)
  decode $result_Q$
  return $result_Q$
end
```

```
match($in\text{-}degree_{vari}$, $out\text{-}degree_{vari}$, $incoming\text{-}edges_{vari}$, $outgoing\text{-}edges_{vari}$)
begin
  access the CS-index based on the in-degree
  and out-degree of $var_i$
  retrieve the values which match with incoming-edges, outgoing-edges
  return $result_{vari}$
end
```

Figure 6. Pseudo code for query processing

When these four values are obtained, the results are searched in CS-index. The location of CS-index where the result can be exist are easily accessed as the CS-index is sorted in ascending order according to the degree of vertices. After the matched value (vertex id) is obtained, all vertex id need to be translated into the original strings by the dictionary lookups. And the system displays the result to user. In this way, the proposed method could reduce the query response time by avoiding number of join operations.

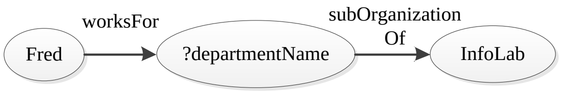
5. Query Evaluation with Proposed Method

Assume that CS-index had been constructed by using CS-index construction algorithm presented in Figure 5. Three queries are considered as case studies for query evaluation using CS-index in Table 2.

The first query,Q_1 retrieves department's name where 'James and Fred' works at and is a sub organization of 'Info Lab'. SPARQL query and query graph pattern of Q_1 are described in Figure 7. Q_1 is a type of chain query and contains only subject-object join.

```
SELECT ?department
WHERE {ub:James rdf:worksFor ?department.
       ?department rdf:subOrganizationOf owl:InfoLab. }
```

(a) SPARQL query



(b) Query graph pattern

Figure 7. Chain query Q_1

Query processing is performed with query processing algorithm described in Figure 6. During query processing, the triple patterns of Q_1 are grouped based on the common join variable. Here, the query Q_1 includes one variable '?department'. And then out-degree, in-degree, outgoing-edges, and incoming-edges values of Q_1 are computed. The values of four parameters are obtained as shown in Table 3.

Table 3. Four parameters' values for Q_1

variable	out-degree	in-degree	outgoing-edges	incoming-edges
?department	1	1	{(6,11)}	{(15,5)}

These values match with line #12 of CS-index (Table 2). Thus, the query processor gets vertex id '10' as the result of Q_1.

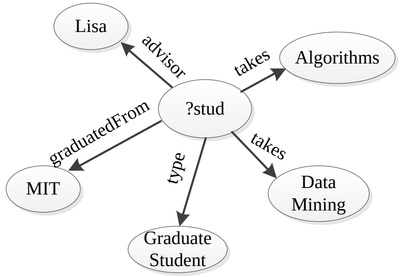
	outgoing-edges (e_i,v_j)	incoming-edges (v_j, e_i)	V_{id}
#12	{(6,11),(3,12)}	{(9,5),(15,5)}	10

This vertex id needs to decode to obtain the original string value. After decoding the vertex id in subject/object dictionary, the query processor returns the result of Q_1 - "CS" department. The query Q_1 requires 2 joins in relation-based RDF stores, but by working with our proposed method, the final result could be get with one unit cost.

As another case study, the second query Q_2 want to retrieve name of graduate student who is Professor Lisa's candidate, got their degrees from MIT, takes both Data Mining and Algorithms course. Q_2 is a star query. It contains subject-subject join. Figure 8 shows Q_2's SPARQL query and query graph pattern.

```
SELECT ?stud
WHERE { ?stud rdfs:type ub:GraduateStudent.
    ?stud rdfs:graduatedFrom "MIT".
    ?stud rdfs:advisor ub:Lisa.
    ?stud rdfs:takes ub:Algorithms.
    ?stud rdfs:takes ub:DataMining. }
```

(a) SPAQL query of Q_2



(b) Query graph pattern of Q_2

Figure 8. Star query Q_2

Query processing of Q_2 is processed as in Q_1. Four parameters' values of Q_2 are obtained as described in Table 4.

Table 4. Four parameters' values for Q_2

variable	out-degree	in-degree	outgoing-edges	incoming-edges
?stud	5	0	{(3,7),(1,2),(2,3),(4,6), (4,5)}	-

The result of Q_2 is found at line #16 in CS-index (Table 2). Vertex id "v_1" is obtained.

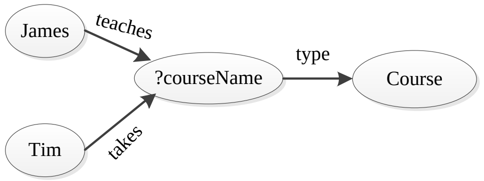
	outgoing-edges (e_i,v_j)	incoming-edges (v_j, e_i)	V_{id}
#16	{(1,2),(2,3),(4,5),(4,6),(3,7)}	-	v_1

To get the original string value, the vertex id "1" is mapped with subject/object dictionary. Finally, the query processor displays the name of student – "Peter". Q_2 need to process 4 joins in relation-based RDF stores. But by employing with our proposed method, the final result could be get with one unit cost.

Next, another star query with subject-subject join and subject-object join is considered. Third query, Q_3 finds course's name Jim studies and is taught by James.

```
SELECT ?courseName
WHERE{ ub:James rdf:teaches ?courseName.
         ?courseName rdfs:type "Course".
         ub:Tim rdf:takes ?courseName. }
```

(a) SPARQL query



(b) Query graph pattern

Figure 9. Star query Q_3 with subject-subject join and subject-object join

Four parameters of Q_2 are computed as in Table 5.

Table 5. Four parameters' values for Q_3

variable	out-degree	in-degree	outgoing-edges	incoming-edges
?courseName	1	2	{(3,13)}	{(9,8), (14,4)}

The result of Q_3 are found in line #11 of CS-index (Table 2). The vertex id obtained from CS-index require to decode as in Q_1. The course's name "Algorithms" is obtained as the query answer of Q_3.

	outgoing-edges (e_i,v_j)	incoming-edges (v_j, e_i)	V_{id}
#11	{(3,13)}	{(1,4),(9,8),(14,4)}	6

Q_3's result is processed in one unit cost even if it contains both subject-subject join and subject-object join.

According to our analysis, the queries with n triple patterns always need to perform (n-1) joins in relation-based RDF stores. The proposed method could minimize the query execution time by avoiding join operations. And it also needs less storage space as it stores CS-index and two dictionaries instead of original RDF data.

6. Conclusion

The proposed approach is designed to gain high-performance query processing for complex SPARQL queries. Typically, when a query with n triple patterns is processed on relation-based RDF stores, (n-1) joins need to perform to get the query's result. It takes too much time for query processing. So, our proposed CS-index and querying approach intend to minimize query processing time by avoiding join operations. The proposed method has index construction time, but it requires only one unit cost to get the result as explained in the query evaluation in Section 5. And it uses reasonable memory space as two dictionaries (subject/object, predicate) and CS-index are needed to store instead of original data set.

In future work, we will prove that our proposed method could efficiently minimize query execution time and memory usage for storing data.

7. References

[1] M.T. Ozsu, "A survey of RDF data management systems", Frontiers of Computer Science, Vol. 10, No. 3, June 2016, pp. 418-432.

[2] S. Sakr, G. AI-Naymat. "Graph indexing and querying: a review", International Journal of Web Information Systems, Vol. 6, No. 2, June 2010, pp. 101-120.

[3] Y. Luo, F. Picalausa, G.H. Fletcher, J. Hidders, and S. Vansummeren, "Storing and indexing massive RDF datasets", In Semantic Search over the Web, Springer Berlin Heidelberg, 2012, pp. 31-60.

[4] T. Neumann and G. Weikum, "RDF-3X: a RISC-style engine for RDF," In Proc. VLDB, pp. 647–659, 2008.

[5] C. Weiss, P. Karras, and A. Bernstein, "Hexastore: sextuple indexing for semantic web data management," In
Proc. VLDB, pp. 1008–1019, 2008

[6] X. Wang, S. Wang, P. Du, and Z. Feng, "CHex: An Efficient RDF Storage and Indexing Scheme for Column-Oriented Databases", International Journal of Modern Education and Computer Science, Vol. 3, No. 3, June 2011, p. 55.

[7] X. Lyu, X. Wang, Y.F. Li, Z. Feng, and J. Wang, "GraSS: An efficient method for RDF subgraph matching", In International Conference on Web Information Systems Engineering, 1 Nov 2011, pp. (108-122), Springer, Cham.

[8] K. Kim, B. Moon, and H.J.Kim, "RP-Filter: A path-based triple filtering method for efficient SPARQL query processing", In Joint International Semantic Technology Conference, 4 Dec 2011, pp. 33-47

[9] K. Kim, B. Moon, and H. J. Kim, "RG-index: An RDF graph index for efficient SPARQL query processing", Expert Systems with Applications, Vol. 41, August 2014, pp. 4596 – 4607.

[10] H. He, A.K. Singh, "Graphs-at-a-time: Query Language and Access Methods for Graph Databases", Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, June 2008, pp. 405-418.

[11] Y. Guo, Z. Pan, and J. Heflin, "LUBM: A benchmark for OWL knowledge base systems", Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 3, 31 Oct 2005, pp. 158-182.

Retrieving Relevant Web Documents by Using a Synonym Based Approach in Mining Web Content Outliers

Thinzar Tun
University of Information Technology,
Yangon, Myanmar
thinzartun@uit.edu.mm,
thinzartunucsy@gmail.com

Khin Mo Mo Tun
Faculty of Computing
University of InformationTechnology,
Yangon, Myanmar
khinmomotun@uit.edu.mm

Abstract

*Due to its convenience and the richnessof information on the web, searching on the web is increasingly becomingthe dominant information seeking method. A large number of users are interested in searching and retrieving information on web. But searching information on theweb,the required information of user contains many irrelevant information. Mining Web Content Outlier concentrates on mining the irrelevant web pages from the rest of the web pages under the same categories.Removal of outliers from webs not only leads to reduction in indexing space and time complexity, but also improves the accuracy of search result.Although the existing approach,the Term Frequency*Inverse Document Frequency (TF.IDF) from information retrieval is used in removing irrelevant web documents return better result, there is need to develop a technique to mine outliers in documents with more precision. And it did not consider the synonyms of each other between terms in documents. This may lead to less search effectiveness, degrades accuracy of search result and mines web content outlier with less precision. In this paper, Term Frequency*Inverse Document Frequency (TF.IDF) technique with synonym based approachis used to remove the irrelevant documents.The dataset is from 20 newsgroup dataset.WordNet is used to take synonyms of the terms in documents.TF.IDF with synonym based approach improves effectiveness and reliability of web search and can detect web content outliers with more precision.*

Keywords- web content outliers, term frequency, inverse document frequency, synonym based approach

1. Introduction

Today, most of the knowledge transfer and business is done through internet as World Wide Web contains voluminous amount of information. With the exponential growth of information available on the web, updating incoming data and retrieving relevant information from the web quickly and efficiently becomes a growing concern. Most web search engines typically employ conventional information retrieval and data mining techniques to automatically discover useful and previously unknown information from the web content. In addition, as most of the data in the web is semi structured and unstructured and contains a mix of text, video, audio, image etc., there is a need to mine information to cater to the specific needs of the users. Efforts are being made to be such data available, usually in some structured form such as table, for querying and further manipulation. The aforementioned problems result in the development of web content mining which uses the ideas and principles of data mining and knowledge discovery to screen more specific data [7].

Web mining has been described as the discovery and analysis of interesting and useful patterns from the web. Web Mining has adapted techniques from the field of data mining, databases and information retrieval. In general, web mining tasks can be classified into three major categories: web structure mining, web usage mining and web content mining. Web structure mining is the discovery of interesting patterns from the hyperlink structure of the web. Web usage mining mines secondary information extracted from user interactions with the web while surfing. Web content mining aims to extract useful information from the web pages based on their contents. So similar pages can be grouped together to enhance performance. Web content mining aim at summarizing information on web pages to facilitate efficient and effective information retrieval [2][5].

2. Background Theory

2.1 Web Content Outlier Mining

Traditional outlier mining aimed at finding rare andinteresting patterns from numeric datasets has receivedtremendous attention recently. However, web outliermining targeting web datasets has not received similarattention in the mining community.

There are threetypes of web outliers; web content outliers, webstructure outliers, and web usage outliers. The problem of identifyingoutliers from web contents called web content outliermining.A web content outlier is described as a webdocument with different contents compared to similardocuments taken from the same category.For example, a loan facility found on an insurance company's website constitutes a web content outlier because insurance companies generally do not grant loan to their customers. The pages dedicated to the promotion of the loan facility will be different from all other pages on the company's website. This paper focuses on designing algorithms to track such web pages with the rare contents [5].

2.2 Information Retrieval and Web Search

Information retrieval (IR) is the study of helping users to find informationthat matches their information needs. Technically, IR studies the acquisition,organization, storage, retrieval, and distribution of information. Historically,IR is about document retrieval, emphasizing document as the basic unit.

Web search has its root in information retrieval (IR), afield of study that helps the user find needed information from a largecollection of text documents. Traditional IR assumes that the basicinformation unit is a document, and a large collection of documents isavailable to form the text database. On the Web, the documents are Webpages.

It is safe to say that Web search is the single most important applicationof IR. To a great extent, Web search also helped IR. Indeed, thetremendous success of search engines has pushed IR to the center stage.Search is, however, not simply a straightforward application of traditionalIR models. It uses some IR results, but it also has its unique techniques andpresents many new problems for IR research [1].

2.3 WordNet

In 1998, a new lexical database called WordNet wasdeveloped for finding the semantic matching of English words. WordNet is used to manage and navigate the entitycomponent on web page. It represents synsets by means ofconceptual semantic and lexical relationship between words.It classifies English words into numerous groups, such as synonyms,antonyms, hypernyms, hyponyms, meronyms and metonyms [6].The synonyms of WordNet are only used in this paper. The synsets are sets ofsynonyms which gather lexical items having similarsignificances, for example the words "car" and"automobile" grouped in the synset {car, automobile}.

The organization of WordNet through lexicalsignificances instead of using lexemes makes itdifferent from the traditional dictionaries and thesaurus. The other difference which has WordNetcompared to the traditional dictionaries is theseparation of the data into four data bases associatedwith the categories of verbs, nouns, adjectives andadverbs. This choice of organization is justified bypsycholinguistics research on the association of wordsto the syntactic categories by humans. Each database isdifferently organized than the others. The names areorganized in hierarchy, the verbs by relations, theadjectives and the adverbs by N-dimensionhyperspaces.

A synonym is a word which we can substitute toanother without important change of meaning. The synonyms can be distinguished into absolute synonyms, cognitive synonymsand plesionyms.Absolute synonymy means, that a pair of lexemes is absolutely interchangeable in all imaginable contexts and that they have the same ratio of distribution.Cognitive synonymy is a type of synonymy in which synonyms are so similar in meaning that they cannot be differentiated either denotatively or connotatively, that is, not even by mental associations, connotations, emotive responses, and poetic value. It is a more precisetechnical definition of synonymy, specifically for theoretical purposes. In usage employing this definition, synonyms with greater differences are often called near-synonyms rather than synonyms. Plesionyms or near synonyms are words, that are almost synonyms. They are distinguished from cognitive synonyms by the fact, that the connotations of the pairs are different and therefore they yield different truth conditions in a given context.

The relation of synonymy is at the base of thestructure of WordNet. The lexemes are gathered in sets of synonyms (synsets). There are thus in a synset allthe terms used to indicate the concept.The definition of synonymy used in WordNetisas follows: "Two expressions are synonymous in alinguistic context C if the substitution of for the otherout of C does not modify the value of truth of thesentence in which substitution is made" [10].Example of the synset: [Person, individual, someone, somebody, mortal, human].

2.4. TF.IDF

The Term Frequency*Inverse Document Frequency (TF.IDF) is a weighting method often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word for a document in a collection or corpus. The TF.IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Variations of the TF.IDF weighting method are often used by search engines as a central tool in scoring and ranking a document's relevance given a user's input query.Typically, the TF.IDF is composed by two terms: the first computes the normalized Term Frequency (TF), the number of times a word appears in a document. Thesecond term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears[11].

3. Related Works

In paper [5], the Relative Document Weight (RDW) concept is used. The term weight assigned to the text in web content depends on which HTML tags enclosed in the text. META and TITLE tags are given a larger weight than BODY tags because it gives a better representation of web content.Most of the web pages that do not have META tag description although the difference documents having varying size within the same category can be compared in RDW.And transforming into the base word is not used in RDW. So it cannot match the various forms of same base word as similar.

The authors modified above technique and use an n-gram method with domain dictionary and without domain dictionary in [3] and [4] to determine the similarity of strings and expand it to include pages containing similar strings. The experimental results show that finding outliers with high order n-grams (5-grams) perform better than lower order n-grams. The existing approach using WCOND Mine algorithm based on n-grams that works only for structured documents. In n-grams,the fixed lengths concept helps in memory utilization and supports partial matching of strings which is good for outlier detection. But n-gram based systems become slow for very large datasets because of the huge number of n-gram vectors generated during mining web content outliers. And the n-gram based system takes a longer time to complete a task than the word-based systems even though the size of data is not too large.

A traditional weighting technique TF.IDF from information retrieval is only compatible to use in detecting web outliers[8]. AlsoTF.IDF with domain dictionary is used in paper [9]. The word-based techniques just maintain the size of the words. Although the words are in variable length, the efficiency of word based web content outlier mining can be increased by indexing the words in two-dimensional format (i, j) and indexing the domain dictionary based on the length of words. The TF.IDF from information retrieval is not only compatible to use in mining web content outliers although it even returns better results than previous works. And they did not consider the synonym of each other between terms in documents. This leads to mine web content outliers with less precision.

In this approach, Term Frequency*Inverse Document Frequency (TF.IDF)with synonym based approach to mine and remove irrelevant web pages from relevant web pages. It improves the web search effectiveness and can mine web outliers with more precision.

4. Architecture Design

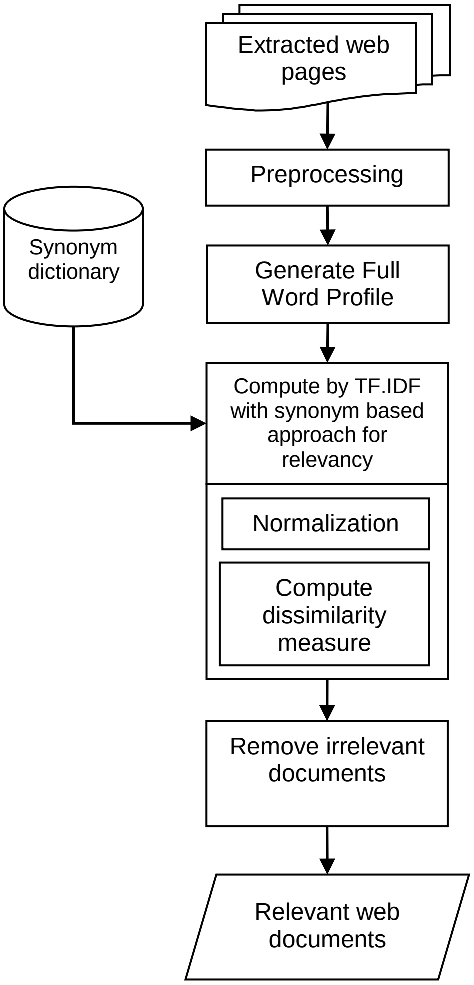

Extracted web pages
Preprocessing
Synonym dictionary
Generate Full Word Profile
Compute by TF.IDF with synonym based approach for relevancy
Normalization
Compute dissimilarity measure
Remove irrelevant documents
Relevant web documents

Figure 1. Architecture design of system

4.1. Web Pages Extraction

The document extraction is the process of retrieving the desiredpages belonging to the category of interest. In this process, the user firstenters a query into a search engine. Based on the keywords extracted fromuser query, the search engine examines its index and provides a list of web pages according to its category. Most of the documents retrieved from search engine may or may not be relevant to the user query.

4.2. Preprocessing

The extracted documents undergo the preprocessing step which consists of stop words removal, stemming and tokenization. Preprocessing is necessary to make the entire document in the same format.

4.2.1 Stop WordsRemoval

Stop words list typically consists of those word classes known to convey little substantive meaning such as articles (a, the), conjunctions (and, but), interjections (oh, but), prepositions (in, over), pronouns (he, it) and forms of the "to be" verb (is, are).Removing stop wordsincreases the efficiency and effectiveness of web search.

4.2.2 Stemming

Stemmingis often used to normalize the morphological variants of the same base word.Stemming removes word suffixes which reduce the number of unique words in the index by reducing the storage space required for the index and speeds up the search process.

4.2.3 Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. A token is a string of characters, categorized according to the rules as a symbol. The list of tokens becomes input for further processing.

4.3. Full Word Profile Generation

After the preprocessing step, the filtered data are then used to generate the full word profile.The words are indexed by using hashing function and stored in a hash table. The hash function isused to map the search key to the hash. The index gives the place where the corresponding record should bestored. Hash function will take an integer key and turn it into an index. In general, a hashing function may mapseveral keys to the same index.Therefore, it is desirable to minimize the occurrence of such collisions bymapping the keys to hash values as evenly as possible [7].

4.4. Computation by TF.IDF with Synonym Based Approach for Relevancy

In this stage, a classic term weighting technique, Term frequency*inverse document frequency (TF.IDF)with synonym based approach isused to evaluate the representativeness of terms in the web content. The dissimilarity measure computed to determine the difference among pages within the same category. The Maximum Frequency Normalization applied to Term Frequency (TF) weighting because the relative frequency is preferredwhen the document length varies. Since term frequency alone may not have the discriminating power to pick up all relevant documents from other irrelevant documents, IDF (Inverse Document Frequency) factor which takes the collection distribution into account has been proposed to help to improve the performance of IR.But it is not effective because it does not consider the synonym of each other between words in documents.

Many concepts or objects can be described in multiple ways using differentterms in web pages due to the context and people's language habits. Some synonyms of the terms (different terms) are used in web pages[1].The IDF factor will becomputed as they aredifferent termsand IDF value will be high discriminative although the termsmay be thesame meaning. For example,the terms"automobile" and"motorcar" are synonyms of "car". i.e. theweb pageuses the term "car", but relevant web documentsthat contain "motorcar" or "automobile", the TDF will not be calculated accurately.

So Inverse Document Frequency with synonym based approach is used instead of Inverse Document Frequency (IDF). And synonym dictionary, WordNet is used to take the synonyms of terms in documents. By using this approach, the IDF value can be computed with precision and can reduce unique terms in IDF. The new approach is:

TF.IDF with synonym based(SB) approach =
(TF) * (Synonym based -IDF)

The dissimilarity equation is below:

$$DM_i = \sum_{i,j} [(0.5 + \frac{0.5 * f(t_j, d_i)}{MaxFreq(d_i)})(log_{10} \frac{N}{k})]$$

where f (t_j, d_i) denotes the frequency of term t_j present in the document d_i, while MaxFreq(d_i) determine maximum frequency of a word in a document, N is the total number of documents and k is the number of documents with term t_j appears [9]. In this approach, the synonyms of term t_j appears in documents is considered. And the IDF factor calculated as they are same terms if the terms in documents are synonyms.

4.5. Determination Relevant Documents

The output from the dissimilarity measure was ranked to determine web content outliers or irrelevant documents. The documents at the top will have high dissimilarity measures deviate more from the category of user interest. The top 'n' documents that have high dissimilarity measure are declared as web content outliers based on threshold value.Also, the documents at the bottom will have less dissimilarity measure which is more relevant to the category of interest and declared as relevant documents.

4.6. Mining Web Content Outliers Algorithm

```
Input: Web Documents $d_i$
Method: TF.IDF with synonym based approach
Output: Relevant web documents
1. Extract the set of documents
2. Preprocess the entire extracted documents by
 stop words removal, stemming and tokenization
3. Generate full word profile
4. For (int i=0; i<NoOfDoc; i++) {
5. For (int j=1; j<=NoOfWords; j++) {
6. Compute dissimilarity measure ($DM_i$)byIF.IDF with synonym based approach
        $DM_i = \sum_{i,j} [(0.5 + \frac{0.5 * f(t_j, d_i)}{MaxFreq(d_i)})(log_{10} \frac{N}{k})]$
7.}}
8. $DM_i$= $DM_i$/ number of words in the document
9. Rank the result of $DM_i$
10. Determine irrelevant documents, remove it and return relevant documents
```

5. Conclusion

The enormous growth of information is available on World Wide Web forcesto develop an effective algorithmfor retrieving relevant documentswith precision.The existing algorithms formining web content outlier cannot consider the synonyms of each other between terms in documents. This lead to mine web content outliers with less precision and less search effectiveness. In this paper, the term weighting technique TF.IDF with synonym based approachis used to remove irrelevant web documents. It can mine web content outliers with more precision and can obtain the accurate results. And it improves the effectiveness of web search.

6. References

[1] B. Liu, "Web Data Mining Exploring Hyperlinks, Contents,and Usage Data", 2nd Edition, Department of Computer Science, University of Illinois, Chicago, ISBN 978-3-642-19459-7, Springer Heidelberg Dordrecht London New York, 2011.

[2] K.Sarukesi, P.Sudhakar, S. Poonkuzhali, "Signed-With-Weight Technique for Mining Web Content Outliers", Special Issue of International Journal of Computer Applications (0975 – 8887) the International Conference on Communication, Computing and Information Technology (ICCCMIT) 2012.

[3] M. Agyemang, K. Barker, and R.S. Alhajj, "Mining web content outliers using structure oriented weighting techniques and n-grams," Proceedings of ACM SAC, New Mexico, 2005.

[4] M. Agyemang, K. Barker, and R.S. Alhajj, "WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents," Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC), 2005.

[5] M. Agyemang, K. Barker, and R.S. Alhajj, "Framework for Mining Web Content Outliers", 2004 ACM Symposium on Applied Computing.

[6] N.K. Jha1, A. Jethva, N. Parmar, A. Patil, "A Review Paper on Deep Web Data Extraction using WordNet", International Research Journal of Engineering and Technology (IRJET),Volume: 03 Issue: 03, March 2016.

[7] P. Gnanasambandan and S. Poonkuzhali, "Proportionate Approach for Retrieving RelevantWeb Documents by using Outlier DetectionMethod", International Journal of Pure and Applied Mathematics, Volume 118 No. 18 ,2018.

[8] W.R.W. Zulkifeli, N. Mustapha, A. Mustapha, "Classic Term Weighting Technique for Mining Web Content Outliers", International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2012). Penang, Malaysia, 2012.

[9] Thinzar Tun, Khin Mo Mo Tun, "Mining Web Content Outliers by using Term Weighting Technique and Rank Correlation Coefficient Approach", 1st International Conference on Advanced Information Technologies (ICAIT), 1st November 2017.

[10] Z.Elberrichi, A.Rahmoun, and M.A.Bentaalah, "Using WordNet for Text Categorization", The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008.

[11] http://www.tfidf.com/TF.IDF

Distributed Computing

ICS:BlockOpS – Blockchain for Operational Data Security in Industrial Control System

Biography
Aung Maw is a graduating student from Yangon Technological University. He completed his final year project at iTrust, Centre for Research in Cyber Security, of Singapore University of Technology and Design (SUTD), Singapore. He is interested in cyber-physical systems security, data security, blockchain beyond cryptocurrency, Industrial 4.0, machine learning.

ICS:BlockOpS – Blockchain for Operational Data Security in Industrial Control System

Industrial Control Systems (ICS) are the backbone of critical infrastructure found in power, water, manufacturing and other industries. An ICS controls a physical plant using sensors and actuators. A Historian sits on a plant network and receives, parses, and saves the data and commands transmitted over the network, across the Programmable Logic Controllers (PLCs), sensors and actuators. The data has at least two uses. One is to check for any process anomalies that may occur due to component failures and cyber-attacks. The other use, and the focus of this presentation, is to serve as critical input to offline activities such as forensic analysis. A cyber-attack on the Historian could jeopardize any forensic analysis be it for maintenance or discovering an attack trail. ICS-BlockOpS, is designed to secure plant operational data recorded in the Historian. A prototype implementation of ICS-BlockOpS uses Ethereum blockchain in the local network as part of the tamper-proofing mechanism. The implementation is in an operational 6-stage water treatment plant. The underlying design ideas are generic and could be applied to other ICS as well.

A Blockchain Based Data Storing on HDFS

Khin Su Su Wai, Ei Chaw Htoon
University of Information Technology, Yangon
khinsusuwai@uit.edu.mm, eichawhtoon@uit.edu.mm

Abstract

A blockchain is a distributed database of record or public ledger of transactions. It is decentralized and shared among the authorized set of users. It has a network of computers which maintain and validate transactions via consensus with cryptographic audit trails. The main idea of blockchain is immutability, traceability, reliable, transparent, anonymity, data integrity and trustworthy between two parties without any trusted third party. However, all transactions of the other nodes are stored in each node. So, much amount of data storage and scalability are not efficiently supported by blockchain. When the implementation scales upwards, the scalability problem is the most serious. If the block size is increased, it also increases the security risks. In order to solve this issue in a blockchain, a new connector approach is proposed to integrate with Hadoop Distributed File System (HDFS). In this paper, the meta-data is only stored on blockchain and the real data transaction is supported on Data Nodes to have more available storage of blockchain based on Message Oriented Middleware (MOM). The limited storage and scalability issue of blockchain infrastructure will be solved by the proposed system.

Keywords- Blockchain, Distributed ledger, Consensus, Cryptographic audit trails, Message Oriented Middleware (MOM), Hadoop Distributed File System (HDFS), DataNode, meta-data

1. Introduction

The blockchain is a structured list that saves data in a form similar to a distributed database. It is designed to make arbitrarily manipulating it difficult since the network participants save and verify the blockchain. Each block is a structure consisting of a header and a body. The header includes the hash values of the previous and current blocks and nonce. The block data are searched in the database using the index method.

The basic blockchain processing consists of the following steps. Firstly, add new and undeletable transactions and organize them into blocks. Secondly, cryptographically verify each transaction in the block. Finally, append the new block to the end of the existing immutable blockchain.

The blockchain is also a peer-to-peer program. It also ensures that all the peers have the same exact data. If the data changes on one machine, it changes on all the machines. There are rules specifying exactly how a change can be made. If someone doesn’t follow them and modifies their copy illegally, they’re ignored. In other words, data is only written, never deleted. The new data is added in blocks and appended to the existing blocks, forming a chain of blocks. Everyone has the same blockchain (database). They get a locker within the blockchain that only they can access. The blockchain has no central authority to manage usernames and passwords. So, it uses cryptography to support privacy and security.

Each user is able to generate a locker address and a private key code that allows them to unlock the locker. The locker is only an analogy. It is just an ID number that referred to as an address, which is tagged to a user’s data. The private key is a code that allows the user to prove they are the creator or owner of that address. Only the person who generated the address would have the private key. No one can ever determine the private key from the address alone.

The hash values stored in each peer are affected by the values of the previous blocks. It is very difficult to falsify and alter the registered data. Although data alteration is possible if 51% of peers are hacked at the same time, the attack scenario is really very difficult. Public, key-based verification and a hash function that can be decrypted are both used to provide security in the blockchain.

Each block size average is 1 MB and contains control data of approximately 200 bytes in public blockchain. Block contains previous block hash, current block hash, timestamp, nonce, and 1 to N transactions as can fit in the remaining space. The larger the blockchain grows, the larger the requirements on storage, bandwidth and computational power. Scalability is a major issue for a public blockchain.

Middleware is a key for success to integrate blockchain with the rest of architecture. A blockchain infrastructure is its own independent peer-to-peer network without a central backbone. Integration is not trivial due to security and governance requirements. Middleware can be built for connecting different systems with various technologies, standards and communication protocols. In addition, middleware can add augmented intelligence via event correlation and visualization to ensure governance requirements. If the integration middleware is built as a connector once, then it can be re-used all the time without any coding to connect Ethereum with other technologies or to apply integration patterns.

Database-oriented middleware is any middleware that facilitates communications with a database, whether from an application or between databases. Developers typically use database-oriented middleware as a mechanism to extract information from either local or remote databases. The developer may invoke database-oriented middleware to log on to the database, request information, and process the information that has been extracted from the database.

In this paper, the new structure of Message Oriented Middleware is proposed to handle the limited storage of blockchain and support better scalability and performance by integrating with HDFS.

The rest of the paper is organized as follows. Section 2 is an illustration of related works with the proposed topic. Section 3 is the background theory of the proposed system. The proposed system is presented in section 4. Finally, the paper is concluded in section 5.

2. Related Works

J. Bambara and et al. [1] described a particular aspect of blockchain technology and its use cases. This is coupled with a comprehensive treatment for getting started as a designer and developer of blockchain applications. This includes the Ethereum technology stack, code and deployment techniques.

X. Xu and et al. [2] provided rationales to support the architectural decision on whether to employ a decentralized blockchain as opposed to other software solutions, like traditional shared data storage. Additionally, this explored specific implications of using the blockchain as a software connector including design trade-offs regarding quality attributes. However, the mining mechanism increases the communication latency, which might cause poor user experience. Likewise, the amount of data that can be stored on the blockchain is very limited. Thus, it is important to decide which data or meta-data should be stored on-chain or off-chain.

E. Ademi [3] took a look at the current research on the challenges and limitations of blockchain, and how these challenges affect the user. Recommendations on future research directions are provided for researchers. There are a number of efforts to tackle the problem of scalability. The proposed idea is to optimize the storage of the blockchain. The blocks are decoupled and there are leader blocks and micro blocks that handle the transactions. Miners would compete for the leader block that would be the ones in charge of the generation of new micro blocks. Increasing the block size or increasing the rate of creation of blocks reduces the security threshold. All of these proposals need to deploy through a hard fork. This means that the new blocks that have a bigger capacity of 1MB will be seen as invalid by the current version node. Segregated witness approach does not increase the block size limit. But it increases the number of transactions that can be stored in a block. This approach in the best case scenario increases the throughput by four times resolves the transaction malleability problem. This allows new mechanisms to be implemented which could provide powerful tools for the scalability issues in blockchain implementations. A variant of GHOST is implemented into Ethereum although the performance has not been tested enough.

D. Puthal and et al. [4] presented a comprehensive review of the blockchain by highlighting the working model of blockchain. This subsequently presented the system features. Consensus algorithms are described with different applications and use cases. Finally, this is concluded with presenting different security challenges.

M. Dai and et al. [5] proposed NC-DS framework to store the blockchain and proposed corresponding solutions to apply NC-DS to blockchain systems. They propose two solutions that deterministic rate (NC-DRDS) and rateless (NC-RLDS) to tackle this problem. Analysis showed that this proposed scheme achieves significant improvement in saving storage room. Trivial application of NC-DS to blockchain is difficult because the parameter is difficult to set.

D.S. Kumar and et al. [6] proposed a simplified architecture for big data storage. This eliminates the concept of master node called Name Node with the functionalities of the Name Node being distributed using blockchain technology. Metadata creation and blockchain placement were implemented and tested in a cluster of nodes. The implementation of blockchain in the proposed methodology results in low metadata access delay and improves the execution time. SHDFS needs further research on the scalability of metadata in blockchain as the cluster capacity was limited. The application data and metadata were sharing the same HDFS cluster. When many numbers of small files are to be maintained, the scalability is being one of the challenges in large-scale data processing.

S.K. Sangode and et al. [7] presented newer HDFS architecture for big data storage. This eliminates the concept of Name Node with the functionalities of the Name Node being stores the metadata using blockchain technology. This new architecture removed expensive hardware and cost-effective strategy. The blockchain stores replicas of metadata on a rack by rack basis with one Data Node/rack and cost of maintaining blockchain. There is no point of recovery due to the presence of several replicas of the metadata being spread throughout one Data Node per rack. Hence, the Name node and secondary name node is not essential in this proposed HDFS architecture.

3. Background Theory

3.1. The Blockchain Technology

The blockchain systems can be categorized into three types: private blockchain, consortium blockchain, and public blockchain. Private or consortium blockchain is linked to a limited environment such as company, group of companies or one specific value chain.

3.1.1. Public blockchain. It maintains the principle that anyone in the world can access the data. This includes the consensus process to write the data into the public blockchain or to block it.

3.1.2. Consortium blockchain. It is partly private. It operates under the leadership of a group instead of a single entity. Reading the data from blockchain may be allowed to public or restricted to a set of participants.

3.1.3. Private blockchain. It allows an only predefined group of users to write any data to blockchain. The data can be read by public or some restricted users.

The private or consortium blockchain is recommended to be used as a framework for the proposed model. Table 1 shows some result from existing evaluations [8] [9]. Hyperledger fabric has cheaper CPU and network utilization. Hyperledger Fabric is a permission blockchain infrastructure. Hyperledger was established under the Linux Foundation. When considering a permission network, blockchain use case needs to comply with data protection regulations. Hyperledger Fabric smart contracts are called chaincode. Chaincode is software that defines assets and related transactions. Chaincode is invoked when an application needs to interact with the ledger. Chaincode can be written in Golang or Node.js. There are the following steps to write a blockchain application. Firstly, chaincode is needed to write in a supported programming language like Go. Secondly, this chaincode is deployed on Hyperledger Fabric network. Finally, a client application using an SDK has developed.

Table 1. Comparison of existing blockchain

Name	Transaction rate	Transaction fee	Network type	Network Access
Bitcoin	3(avg;), 7(max;)	Yes	Public	Permission less
Ethereum	Depend on eth	Yes	Public (or) Private	Permission less
Hyperledger fabric	>10k	No	Private	Permissioned

The external data cannot be accessed by blockchain on their own network. It will also need to be integrated with the third-party services such as oracles, agents, or data feeds. The real-world occurrences are typically accessed and verified by these oracles or agents. This information is submitted to a blockchain to be used by smart contracts. The external data can be provided when needed and push it onto the blockchain by these oracles or agents. In this proposed system, HDFS is used to provide the blockchain.

3.2. Hadoop Distributed File System (HDFS)

Hadoop is a software framework which is an open source that supports big data in the distributed environment. It has two major components HDFS and Map Reduce. HDFS has master-slave architecture, with a single master called the NameNode and multiple slaves called DataNodes. NameNode manages the metadata and regulates client accesses. The metadata is maintained in the main memory of the NameNode to ensure fast access to the client, on reading/writing requests. DataNodes provide block storage and service read/write requests from clients, and perform block operations by contacting with NameNode.

The consumption of memory in NameNode is decided by the number of files stored in HDFS. Each file requires 150 bytes of memory space to store metadata in NameNode. DataNode is responsible for saving the real and replicated data. Each file is split into several blocks with the size of 128MB. The DataNode keeps on sending the heartbeat signal to the NameNode at regular intervals to indicate its existence in the system. The heartbeat consists of DataNode’s capacity, used space, remaining space, and some other information.

3.3. Message Oriented Middleware (MOM)

Message Oriented Middleware (MOM) is a software/hardware infrastructure that supports the receiving and sending of messages over distributed applications. The spreading of applications over various platforms and the creation of software applications comprising many operating systems and network protocols are made less complicated. It is one of the most widely used types of middleware.

MOM allows various software components to talk to each other or share data. In visual models, this type of middleware is often represented as a central station with lines that connect different technologies involving message origination and delivery destinations. It is at this most critical point that it needs to integrate new components or to scale existing ones as efficiently as possible.

Applications distributed on different network nodes use the application interface to communicate. In addition, a virtual system of interconnected applications can be made reliable and secure by providing an administrative interface. MOM system resolves issues like interoperability, reliability, security, scalability, and performance.

Using a MOM system, a client can make an API call to send a message to a destination managed by the provider. The call invokes provider services to route and delivers the message. Once it has sent the message, the client can continue to do other work, confident that the provider retains the message until a receiving client retrieves it.

Many message-oriented middleware implementations depend on a message queue system. In a message-based middleware system, the message received at the destination need not be identical to the message originally sent. A MOM system with built-in intelligence can transform messages route to match the requirements of the sender or of the recipient. Many modern MOM systems provide sophisticated message transformation (or mapping) tools which allow programmers to specify transformation rules applicable to a simple GUI drag-and-drop operation.

4. Proposed System

A blockchain is one of the technologies for the safe distribution of metadata over a distributed system. Blockchain can support a cryptographic approach for identifying and authenticating the users and can also provide an immutable storage. However, the storing large document in blockchain is computationally expensive. It must be replicated across all nodes in the network. Blockchain structures can be integrated with the Hadoop to handle large files and scalability.

The contents of the blockchain are verifiable and immutable. The blockchain technology can be adopted on HDFS for storage of metadata in the blockchain database and real data transaction in Data Nodes. HDFS is available of storage and processing of terabytes or petabytes of data. Metadata contains the information about the block size, distribution of data, replica maintenance. The metadata information is stored in blockchain as encrypted format and a distributive fashion. HDFS has been adopted to support Internet applications because of its reliable, scalable and low-cost storage capability.

Now, Hadoop and blockchain are evolved as effective tools to tackle issues in different domains. Big data applications can help blockchain by managing structured datasets of wallet addresses and their owner details. A blockchain for big data can create a decentralized data resource. This can also track updates to the data resource by eliminating the need and confusion due to multiple copies. In this proposed system, the concept of DataNode will only be used to store the data transaction. The metadata will be stored by using the blockchain technology instead of NameNode. The private blockchain will be used to achieve performance, emergency consumption, and scalability challenges. The proposed system architecture is shown in figure 1.

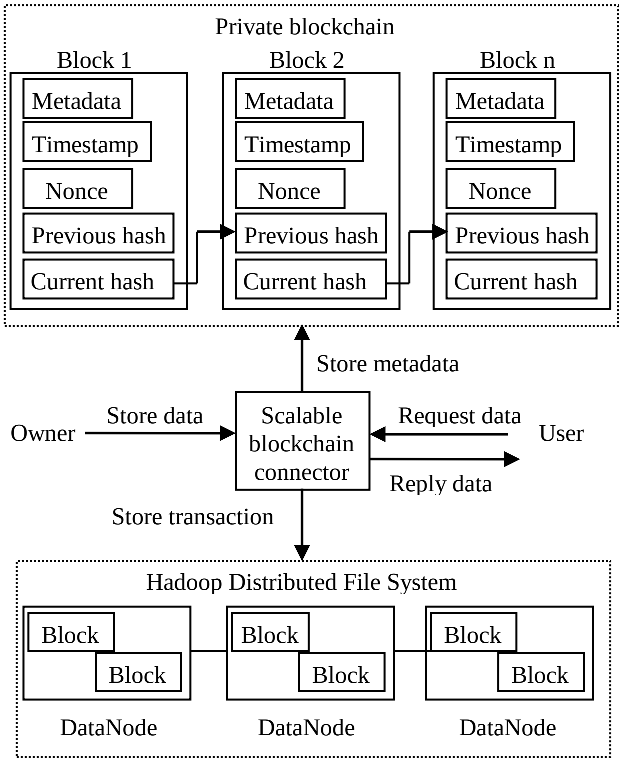


Figure 1. Proposed system architecture

The owner is an entity who owns the secret data and wishes to read or write the data from and to the system. The owner has a full right to control all accesses on his/her own data. He/she is responsible for defining an access policy including connector disciplines. The scalable blockchain connector is responsible for verifying the authenticity relating to access policy, saving the real data transaction on DataNode of HDFS, storing the meta-data on the private blockchain. The HDFS is responsible for storing the real data transaction on DataNode. The private blockchain is responsible for storing meta-data and it can be accessed within a predefined group of users.

5. Conclusion

In this paper, the scalability issue of blockchain is focused by integrating with HDFS based on Message Oriented Middleware. The meta-data will only be saved on blockchain architecture. The DataNode of HDFS architecture will be used to store the data transaction. Furthermore, a new way of the connector will be provided. The limited storage size and scalability of blockchain will be handled by the proposed system. As a future work, many experiments have to be done in order to get the efficiency of the proposed system by using the private blockchain. Moreover, a mechanism will be considered as a future work to retrieve the data from the proposed system.

6. References

[1] J.J. Bambara, P.R. Allen, K. Iyer, S. Lederer, R. Madsen and M. Wuehler. "Blockchain: A practical guide to developing business, law, and technology solutions", 2018.

[2] X. Xu, C. Pautasso, L. Zhu, V. Gramoli, A. Ponomarev, A.B. Tran and S. Chen. "The blockchain as a software connector". In Software Architecture (WICSA), 2016 13th Working IEEE/IFIP Conference on (pp. 182-191).

[3] E. Ademi. "A Comprehensive Study on the Scalability Challenges of the Blockchain Technology", 2018.

[4] D. Puth al, N. Malik, SP. Mohanty and E. Kougianos. "Everything You Wanted to Know About the Blockchain: Its Promise, Components, Processes, and Problems". IEEE Consumer Electronics Magazine 7(4), 6-14, 2018.

[5] M. Dai, S. Zhang, H. Wang, and S. Jin. "A Low Storage Room Requirement Framework for Distributed Ledger in Blockchain". IEEE Access, 6, pp.22970-22975, 2018.

[6] D.S. Kumar and M.A. Rahman. "Simplified HDFS Architecture with Blockchain Distribution of Metadata". International Journal of Applied Engineering Research, 12(21), pp.11374-11382, 2017.

[7] S.K. Sangode and H.K. Barapatre. "Generate Distributed Metadata using Blockchain Technology within HDFS Environment". International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 03 | Mar-2018.

[8] M. Pustišek and A. Kos. "Approaches to Front-End IoT Application Development for the Ethereum Blockchain". Procedia Computer Science, 129, pp.410-419, 2018.

[9] Z. Hintzman. "Comparing Blockchain Implementations". SCTE/ISBE, 303-517-2664, 2017.

[10] https: // en.wikipedia.org/ wiki/ Middleware (distributed applications) - Wikipedia.htm.

[11] https: // en.wikipedia.org/ wiki/ Message- oriented middleware.

Steady State Availability for Hosts in Software Defined Networks

May Thae Naing, Ei Thin Su
University of Information Technology, Yangon, Myanmar
maythae@uit.edu.mm, eithinsu@uit.edu.mm

Abstract

The high-speed and complicated network of hosts and network devices often meet with a variety of failures due to links or system components. This failure affects the availability of the system. The proposed system uses the hierarchical modeling approach that is used to evaluate the availability of SDN infrastructure. This paper offer Steady State Availability (SSA) for host(s) in software defined network (SDN) Infrastructure and then describe the Stochastic Reward Net (SRN) model for the availability of hosts in SDN. Moreover, the impact of software and hardware failures on the overall availability of SDN hosts is evaluated by SHARPE Tool.

Keywords - Software Defined Networks, Availability, Hierarchical Modeling, Steady State Availability, Stochastic Reward Nets.

1. Introduction

Software-Defined Networking (SDN) is an emerging networking paradigm that gives hope to change the limitations of current network infrastructures. First, it breaks the vertical integration by separating the network's control logic (the control plane) from the underlying routers and switches that forward the traffic (the data plane). Second, with the separation of the control and data planes, network switches become simple forwarding devices and the control logic is implemented in a logically centralized controller (or network operating system), simplifying policy enforcement and network (re)configuration and evolution [1], [2].

The key concept of the cloud computing is virtualization. Virtualization is the abstraction of the physical resources needed to complete a request and underlying hardware used to provide service. And also, the idea of the SDN is adopted from the concept of virtualization, where the controls and managements of software subsystems are completely decoupled from hardware infrastructure. The decoupled components of the SDN are separated into three layers of the SDN architecture; (i) Data plane: SDN enabled network devices on a data plane reside at the bottom of the SDN architecture as the underlying physical layer, (ii) Control plane: network operating systems and hypervisors on the control plane resides at the middle layer to provide a bare virtualized environment; and (iii) Management plane: network applications running on the management plane resides at the uppermost layer. This virtualization approach brings three key attributes to the SDN: logically-centralized intelligence, programmability and high-level abstraction. Nevertheless, there are still many issues to use SDNs [3]. In fact, physically centralized network infrastructure still requires adequate levels of system availability and reliability.

High availability refers to the ability of a system to perform its function continuously (without interruption) for a significantly longer period of time than the reliabilities of its individual components would suggest. High availability is most often achieved through fault tolerance. Therefore, the effort in the proposed system will offer an availability model by a comprehensive evaluation of the SDN infrastructure. To evaluate the model using SHARPE tool simulation is presented.

This paper organizes as follows: Section II describes the related work of the proposed system, Section III presents the hierarchical modeling approach, and Section IV describes the steady-state availability of the model. The numerical results and discussion are presented in Section V. Finally, Section VI concludes the paper.

2. Related Work

One of the main reasons of hesitating to adopt SDNs is the concern on availability. There are a few works on the availability of SDNs. In paper [4], the authors considered the impact of SDN application failures on the controller reliability. In paper [5], the authors proposed a stochastic model focusing only on the controller of an SDN rather than the whole SDN. In paper [6], the authors presented experimental results to improve the reliability and availability of core networks using SDN/Openflow. In paper [7], the authors proposed an approach to provide high-availability applications using an SDN. In paper [8], the authors proposed a stochastic model focusing a stochastic availability model with the incorporation of hardware failures and software failures. They used RAID1 architecture for storage system.

In paper [9], the authors formalized a two-level availability model that is able to capture the global network connectivity without neglecting the essential details. It has highlighted the considerable impact of operational and management (O&M) failures on the overall availability of SDN. Moreover, its results showed that the impact of software and hardware failures on the overall availability of SDN can be significantly reduced through proper over provisioning of the SDN controller(s). The paper [10] provided similar availability to the traditional IP backbone networks. It also used a two-level availability model which is able to capture the global network connectivity without neglecting the essential details and which includes a failure correlation assessment should be considered. It also presented the implementation on M¨obius of the Stochastic Activity Network (SAN) availability model of the network elements and the principal minimal-cut sets of an SDN backbone network and the corresponding traditional backbone network.

3. Hierarchical Models for the SDN

This section presents the steps to construct a hierarchical availability model for a SDN system, as depicted in Figure. 1. The availability of the SDN is modeled via hierarchical models in which Stochastic Reward Net model is used to represent the failure and recovery behaviors of hosts. The SDN inputs shown in the left-most box are specifications and parameters of network component host(s). By using the inputs, the core procedure is executed as presented in the middle box which consists of two steps (1) construction of Stochastic Reward Net model and (2) availability analysis. The model construction step involves generating the Stochastic Reward Net model from the components' specifications. In this example, the dependability is measured in terms of steady state availability (SSA), in the following referred to as availability.

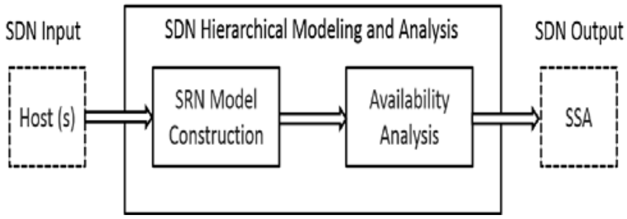


Figure 1. Overview of the proposed system

The approach seeks to avoid the potential uncontrolled growth in model size, by compromising the need for modelling details and at the same time modelling a (very) large scale network. The detailed modelling is necessary to capture the dependencies that exist between network elements and to describe multiple failure modes that might be found in some of the network elements and in the controllers. The structural model disregards this and assumes independence between the components considered, where a component can be either a single network elements with one failure mode or a set of elements that are interdependent and/or experience several failure modes and an advanced recovery strategy. For the former we need to use dynamic models such as a Markov model or Stochastic Petrinet (e.g., Stochastic Reward Network [11]), and for the latter structural models such as reliability block diagram, fault trees, or structure functions based on minimal cut or path sets.

The objective of the modeling approach is to evaluate the availability of SDN.

4. Steady-State Availability of the Model

In this evaluation, this paper considers the network topology depicted in Figure 2. The service provider accesses the SDN controller through the northbound APIs. The service provider uses them for configuring the network resources and adding or removing tenants. The tenants use the northbound APIs to create subnets, to define policies. The controller controls the network switches and gateways through OpenFlow. We distinguish between the core network and the edge. The latter comprises the access switches to which end hosts are connected and the gateways that connect the service provider's network to external networks.

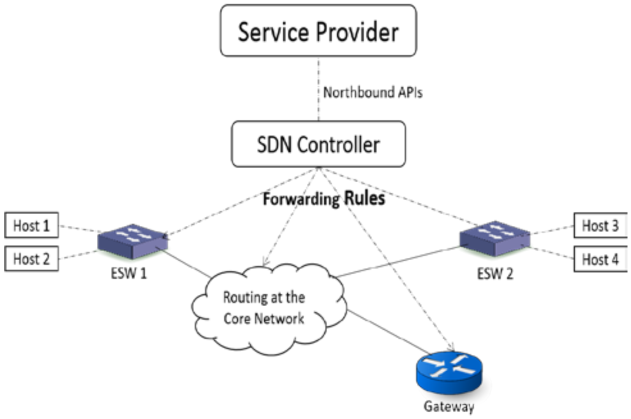


Figure 2. An example SDN Architecture

A host may undergo a failure residing in its hardware components (e.g., CPU, local memory, network interface card etc.) [12]. An unexpected failure may also occur on Virtual Machine Monitor (VMM) as in [13]. Both cases result in a downtime of the host. To simplify the availability modeling, we take into account the failure of either hardware or VMM as a common failure that causes the host go down. To analyze the steady-state availability model, we also use Mean Time to Failure Equivalent (MTTFeq) and Mean Time to Recovery Equivalent (MTTReq) for a host failure and recovery.

5. Stochastic Reward Net Model of the SDN Host

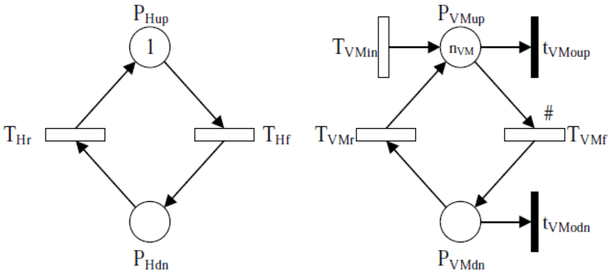


Figure 3. SRN model of a SDN Host

A host's failure and recovery behaviors are simply captured by using a two-state model as shown in Figure. 3. The running and down states of the lower layers underneath a VM (hardware system and VMM) are represented by the places P_{Hup} and P_{Hdn}, respectively. Their failure and recovery are captured through enabling/disabling the timed-transition T_{Hf} and T_{Hr}, respectively. The operational and down state of a VM are depicted by the places P_{VMup} and P_{VMdn}, respectively. The enabling of the T_{VMf} and T_{VMr} respectively represents for failure (with competition between VMs depicted by '#' for rate dependency in which the rate of the failure transition varies upon the number of competing VMs) and recovery of a VM. Initially, there are a number n_{VM} of VMs running on a host. To capture the dependency:

1. As T_{Hf} is enabled and the token in P_{Hup} is removed and deposited into P_{Hdn}, then the immediate transitions t_{VMoup} and t_{VModn} are enabled to remove all tokens in P_{VMup} and P_{VMdn}. Those imediate transitions are used to clear up all VMs on the host.

2. As T_{Hr} is enabled to remove the token in P_{Hdn} and to deposit a token into P_{Hup} (the host is repaired), the timed-transition T_{VMin} is enabled to deposit one after another the tokens into P_{VMup}, which is meant that booting VMs is done successfully.

5.1. Steady State Availability Analysis

The SRN model analyzed using Symbolic Hierarchical Automated Reliability and Performance Evaluator (SHARPE) [14].

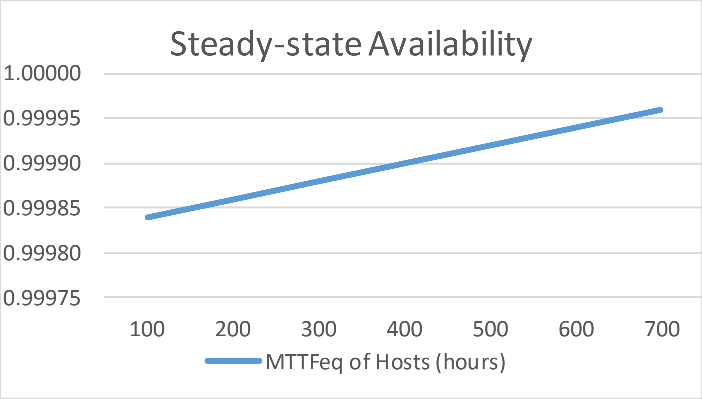


Figure 4. MTTFeq of Hosts vs. SSA

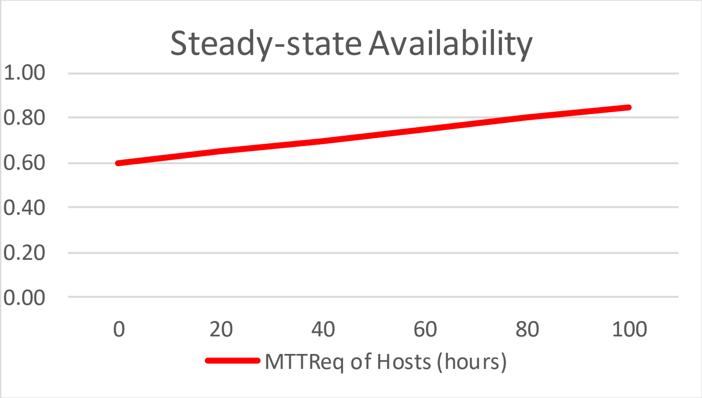


Figure 5. MTTReq of Hosts vs. SSA

6. Conclusions

This paper offer availability solution for Software Defined Network (SDN) Infrastructure. The hierarchical model that includes structural and dynamic models has been formalized and for the dynamic level SRN model of the single network elements have been proposed. The numerical analysis used to evaluate the availability model. In the future, this proposed system will evaluate through both analytical and simulation tool (SHARPE).

7. References

[1] N. Mckeown, "How SDN will Shape Networking," October 2011. [Online]. Available: http://www.youtube .com/ watch?v=c9-K5O qYgA.

[2] S. Schenker, "The Future of Networking, and the Past of Protocols," October 2011. [Online]. Available: http://www.youtube.com/watch?v= YHeyuD89n1Y.

[3] D. Kreutz, F. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," Proceedings of the IEEE, vol. 103, no. 1, pp. 14–76, Jan 2015.

[4] B. Chandrasekaran and T. Benson, “Tolerating SDN Application Failures with LegoSDN,” in Proceedings of the Third Workshop on Hot Topics in Software Defined Networking, ser. HotSDN ’14. New York, NY, USA: ACM, 2014, pp. 235–236. [Online].

[5] F. Longo, S. Distefano, D. Bruneo, and M. Scarpa, “Dependability modeling of Software Defined Networking,” Computer Networks, Apr. 2015.

[6] M. Tamura, T. Nakamura, T. Yamazaki, and Y. Moritani, “A study to Achieve High Reliability and Availability on Core Networks with Network Virtualization,” Technical Report, vol. 15, no. 1, Jul. 2013.

[7] S. Dwarakanathan, L. Bass, and L. Zhu, “Application Level HA and QoS Using SDN,” NICTA Technical Report, Tech. Rep., 2015.

[8] K. Han, T. A. Nguyen, D. Min, and E. M. Choi, “An Evaluation of Availability, Reliability and Power Consumption for a SDN Infrastructure Using Stochastic Reward Net,” Advances in Computer Science and Ubiquitous Computing, vol. 421, 2016, pp 637-648.

[9] G. Nencioni, B. E. Helvik, A. J. Gonzalez, P. E. Heegaard, and A. Kami´nski, “Availability Modelling of Software-Defined Backbone Networks,” submitted to the 2nd Workshop on Dependability Issues on SDN and NFV (DISN 2016).

[10] G. Nencioni, B. E. Helvik, P. E. Heegaard, “Implementing the Availability Model of a Software-Defined Backbone Network in M¨obius,” submitted to the 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2017.

[11] G. Ciardo and K. S. Trivedi, “A decomposition approach for stochastic reward net models,” Perf. Eval, vol. 18, pp. 37–59, 1993.

[12] W. E. Smith, K. S. Trivedi, L. A. Tomek, and J. Ackaret, “Availability analysis of blade server systems,” IBM Systems Journal, vol. 47, no. 4, pp. 621–640, 2008.

[13] D. S. Kim, F. Machida, and K. S. Trivedi, “Availability Modeling and Analysis of a Virtualized System,” in Proceedings of 2009 15th IEEE Pacific Rim International.

[14] K. S. Trivedi, “SHARPE 2002: Symbolic hierarchical automated reliability and performance evaluator,” in Proceedings of the 2002 International Conference on Dependable Systems and Networks. IEEE Comput. Soc, 2002, p. 544.

FRAM Analysis for Continuous Integration Model Implementation

Phyu Phyu Than
High Performance Computing, University of Information Technology
phyuphyuthan@uit.edu.mm

Abstract

In modern enterprise environment, there are many business requirements for software industries. When the business requirements are changing frequently, it also needs to make changes in software development. In the traditional software development, code changes often lead to delay time, loss resources and increase cost for software release. One of the most awkward and tense phases in a software development life cycle (SDLC) is integration step. Continuous integration (CI) is one of the most popular and commonly aspired to agile practices. This paper will analyze a traditional integration model by analyzing with Functional Resonance Analysis Method (FRAM). And a reliable CI model will be proposed. Every components of the whole model will run faster and reliably with efficient manners to improve the quality and efficiency of software that leads to meet the real value of business in continuously and rapidly.

Keywords- Continuous Integration, Software Development Life Cycle, Functional Resonance Analysis Method

1. Introduction

In traditional software production pipeline, software functional modules that worked individually were integrated together and it was very difficult to find out the bugs or errors when the whole program went wrong. Among SDLC models, Agile methodology is an iterative approach to product delivery that builds incrementally from the start of the project, instead of trying to deliver the entire product at once near the end. Continuous integration (CI) is commonly aspired to agile practices Figure 1. Most of software industries have awareness in the automation of software development pipeline that delivers significant benefits to the business. But not all industries are on the way of automation and they don’t have practical benchmark for the level of automation. If the same processes were repeated, then they can be automated. The entire pipeline should be as automated as makes sense for each organization. Whenever a developer commits in every single time, updated result will go to production in a few minutes. Each and every developer don’t need to sync every time. Build automation with one main script and test automation will be in clean environment. Developer team will be notified of build results through a feedback mechanism or report.

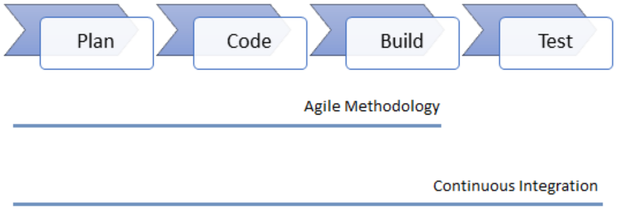


Figure 1. Phases of Agile Methodology and Continuous Integration

2. Background

Continuous Integration (CI) is practical use in software engineering that merges the working copies of all developers to a common repository several times a day. Grady Booch first proposed CI in 1991 and this concept is adopted by extreme programming. CI is the pre-requisite and former step in implementing Continuous Deployment (CD). The essential components in CI are developers or development team, source control server or version control repository, and continuous integration server. The main purpose of CI is automation of sequential steps that the developers commit code changes to repository, the CI server checks and pull the code, build automation through continuous test environment and report every commit through dashboards to provide project insights to developers and managers. CI tends to reply fast when encounter failures and fix the code in time. Proper usage of CI may reduce software failure risks, repetitive manual processes, generate deployable software at any time and any place, produces high quality product that provide better scalability, flexibility, increased project visibility, and finally strengthened the greater confidence in the operational team and development team [1][2]. In previous study, there is no reference about analyzing CI with FRAM. The purpose of the FRAM is to analyze how something has been done, how something is done, or how something could be done in order to produce a representation of it in a reliable and systematic manner, using a well-defined format. This resulting representation is effectively a model of the activity because it captures the essential features of how something is done. In the case of the FRAM, the essential features are the functions that are necessary and sufficient to account for the activity together with the way in which the functions are coupled or mutually dependent [3]. FRAM focuses on variability and safety as system quality that is necessary and sufficient to ensure. FRAM has safety I or safety II method; those will be used to achieve the adverse outcomes as low as possible.

3. Problem Definition

The integration process can be performed both manually and automatically. In the manual integration process, there has lots of human errors where the developers attempt for the clean build. The clean build helps to get ready product with high quality and improves software productivity. Since the integration is a tedious task, as in [2] the manual integration sometimes leads to software failure. In the automated CI process, adopting correct tools is necessary for enabling CI but it can vary on the experiences of DevOps, organization's rules and policies, software infrastructure and business requirements. The lack of knowledge of continuous integration tools and its environment, lack of automation steps, take long time to learn and fix failures, produce lower quality software are going to be problems in CI. Continuous integration model in general is not too complex and affordable for software industry. Some CI tools automate the wrong processes, use unmeaningful metrics and lack of balancing resource utilization for processes. Certainly no one can ensure that CI model is actually available, scalable and reliable.

4. Proposed Idea

The proposed idea will be about two parts. Analysis of traditional integration and implementing a new CI model. Analysis of traditional integration with FRAM can ensure which functions have variability and lead to an accident. FRAM has six aspects in a function. But it isn't necessary to have six aspects in every function or component of integration process. It can define variability going between upstream and downstream functions and go spreads from one function to another. This result can minimize the variability of accident cautions. In second part, the paper will be proposed a new CI model. Automated integration is better than traditional integration. When implementing the Continuous Integration model, the single design is not always perfect for all organizations. It depends on working behaviors of the organization's software development team and sometimes makes changes in communications between CI components. This paper will follow best practices of CI and provide a complete automated CI model that may contain basic CI flows and additional steps for continuous delivery. The automation of code analysis, build, test selection, test scheduling, and test execution process will run for every code change. This paper will be analyzed integration detail with FRAM and supposed a new reliable CI model design. The simple workflow of the proposed idea was shown in Figure 2.

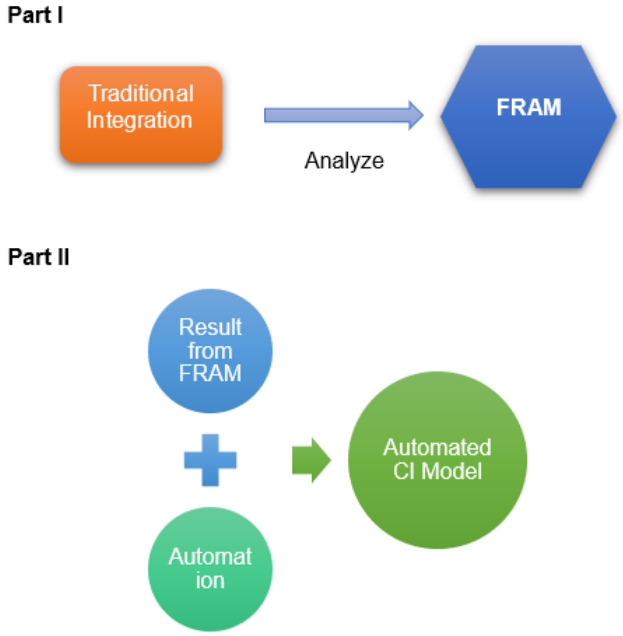


Figure 2. Overview of proposed system

5. Conclusion

The Continuous Integration model will certainly reduce the integration risks by waiting to merge the work of all developers during the development cycle. The build automation will reduce the amount of time that wasted on fixing bugs or errors by each developer. The test automation will help no more effects on the whole program if some regressions occur. The report automation will automatically report to developers and managers who work many projects to get immediately reports for every change in each software product. This model helps to understand conflicts and share better understanding of codebase among developers. It will release bug free high-quality software product continuously. Finally, the software product will be delivered to customers earlier than before. Instead of using traditional integration, using the proposed CI model can suppose significant benefits to organizations.

6. References

[1] Jovanovic, “Software Testing Methods and Techniques”, IPSI BgD Trans, InternetRes, 2009.

[2] Duvall, G.A., Paul, M. and Steve, M., “Continuous Integration: Improving Software Quality and Reducing Risk”, Pearson Education, 2007.

[3] Erik Hollnagel, “The Functional Resonance Analysis Method”, 2018.

[4] Sten Pittet, “How to get started with Continuous Integration”.

[5] Guihuan Duan, Jin Tian, Juyi Wu, “Extended FRAM by Integrating with Model Checking to Effectively Explore Hazard Evolution”, 2015.

Image Processing

Autonomous Lane Boundaries Detection System based on Segments' Angles Computation with Retinex Algorithm

Shwe Yee Win, Htar Htar Lwin
University of Information Technology
shweyeewin@uit.edu.mm,htarhtarlwin@uit.edu.mm

Abstract

Lane detection is important for road security improvement in unmanned vehicle systems. In this paper, an efficient lane detection system to deal with different types of lighting conditions has been developed. Retinex algorithm is employed to normalize input images for all types of illumination. Moreover, Segments' Angles Computation based on Hough Transform is utilized to detect lane boundaries correctly. The experimental results show that the proposed method can reduce computation time and detect left and right boundaries accurately.

Keywords-Lane boundaries detection, Hough Transform, Retinex

1. Introduction

Vision-based lane marking detection and moving vehicle detection are required to the driving safety. In recent years, various approaches were proposed and improved. Detecting lane boundaries enables vehicles to avoid collisions and warn if a vehicle passes a lane boundary. There are two types of lane boundary detection methods: model-based and feature-based methods [6, 8, and 10]. In model-based methods, lane boundaries are presented by mathematical models while feature-based methods use segmentation methods to locate road areas. Moreover, model-based methods usually require a very complex modeling process involving much prior knowledge and background. Among model-based and feature-based methods, feature-based algorithms are efficient and popular.

Fully autonomous vehicles typically use computer vision for producing a map of its environment and for detecting obstacles. The level of autonomy ranges from fully autonomous (unmanned) vehicles to vehicles where computer vision based systems support a driver or a pilot in various situations. To achieve autonomous navigation, the correct recognition of lane detection is the most important issue for automobile vehicles in safety driving assistance system. For autonomous vehicle, detection of lane suffers from high computational complexities. Moreover, if the illumination changes, there is reflection on the road and lane markings are blurred, the lane detection precision may be deteriorated. To solve these problems, in this paper, we propose Segments' Angles Computation based on Hough Transform with Retinex Algorithm for lighting problem.

This paper is organized as follows: In section 2, reviews of researches related to automotive lane detection systems are provided in the literature. In section 3, our proposed method for lane boundaries detection is described for different lighting conditions. Section 4 presents the relevant experiment results. Finally, section 5 describes conclusion and future work.

2. Research Background

The correct recognition of lane is the core issue of safety driving assistance system such as lane departure warning system for intelligent vehicles to achieve self-autonomous navigation. Lane boundaries can be considered as a visual information feedback of the road environment to the drivers. Compared to other measurement methods, visual sensors are low cost and easy to construct an application system, and thus computer vision-based lane detection methods have been widely used to obtain the position information of the lane.

Cheng et al. [1] described a method based on color. They get the information from each color channel of original RGB images, and based on that, extract lane markings respectively. However, this algorithm is easily affected by the light. For example, in the condition of strong light, it will get a wrong detection result.

F. Mariut et.al [2] proposed an algorithm that automatically emphasizes the lane marks and recognizes them from digital images, by the use of Hough transform. This method also detects lane mark's characteristics and has the ability to determine the travelling direction. A technique that extracts the inner margin of the lane is used to ensure the right detection of the lane mark. The algorithm works very efficiently for straight roads but fails in some cases of curved roads.

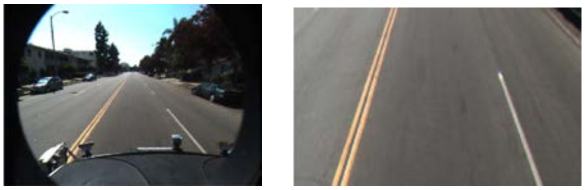
Z. Teng et al. [9] proposed an algorithm which integrated multiple cues, including bar filter which has been efficient to detect bar-shape objects like road lane, color cue, and Hough Transform. To guarantee the robust and real-time lane detection, particle filtering technique has been utilized. This algorithm improved the accuracy of the lane detection in both straight and curved roads. It has been effective on a wide variety of challenging road environments. This method fails for the lane tracking when it is to be applied to particle filter in the dashed lane situation.

Hao Yu et al. [7] described a constraint between Sobel operator and Shen Jun edge operator to detect the lane marking points. In addition to, they presented a new vehicle detection algorithm which uses the shadow under the vehicle and the vertical edge to detect the candidate vehicles. Then they used Support Vector Machine (SVM) and Histogram of Oriented Gradients (HOG) to verify their system.

3. Innovation of Method

Our proposed method for lane boundaries detection is described in Figure.2. In our detection system, the algorithm has four stages: setting region of interest (ROI), preprocessing, reducing lighting conditions and detecting line segments by segments' angles computation based on Hough Transform.

In the first step, we define the ROI of the captured image. Then, in pre-processing step, two sub-stages are described: conversion color image into gray scale image and noise removal. The major purpose of our system is to detect lanes under different lighting conditions and to improve performance and processing time of lane detection system.



(a)Input image (b) ROI Image

Figure 1. Setting region of interest (ROI): (a) original input image; and (b) ROI image

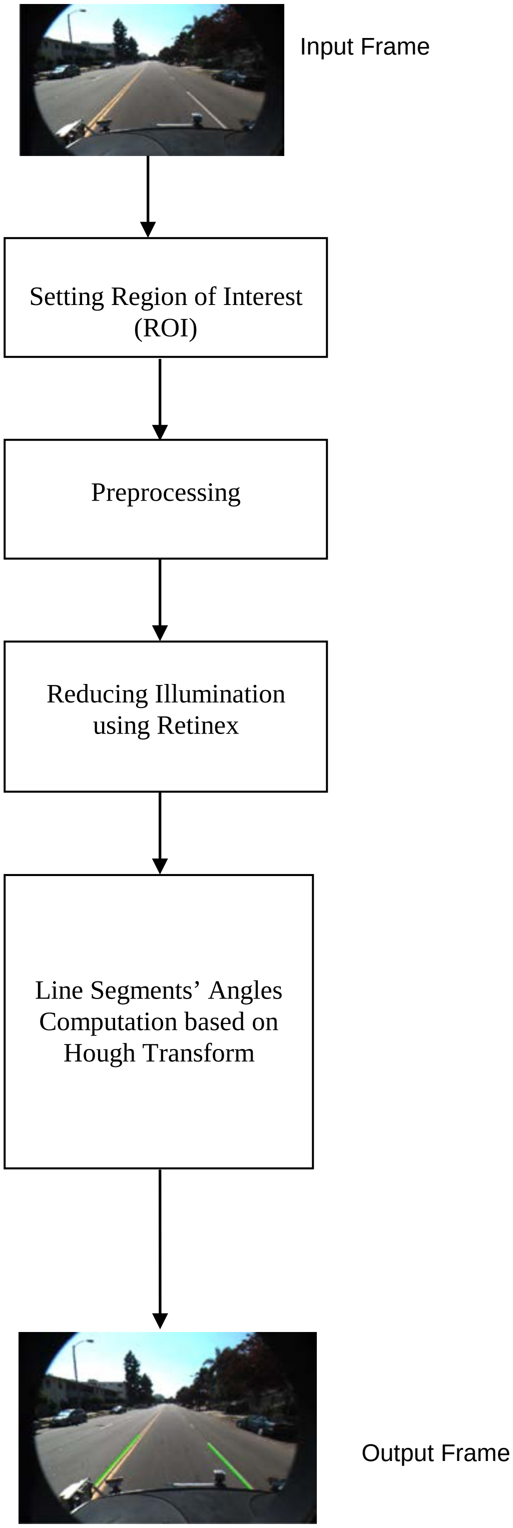

Input Frame
Setting Region of Interest (ROI)
Preprocessing
Reducing Illumination using Retinex
Line Segments' Angles Computation based on Hough Transform
Output Frame

Figure 2. Proposed lane boundaries detection method

3.1. Establishing Region of Interest (ROI)

First, ROI (Region of Interest) of the input image is determined to get only region of image which consists of relevant information (i.e. lanes' boundaries) as shown in Fig.1.It gives us two advantages: reducing complexity and computational time.

3.2. Preprocessing

Next, input color image is converted into grayscale image in order to improve processing speed as shown in Fig.3. This process converts a 24-bit color image to an 8-bit grayscale image because it affects the processing time.

The last step of preprocessing stage is de-noising as depicted in Fig.4. In this case, we apply Gaussian filter to remove noise because of the presence of noise in the image.

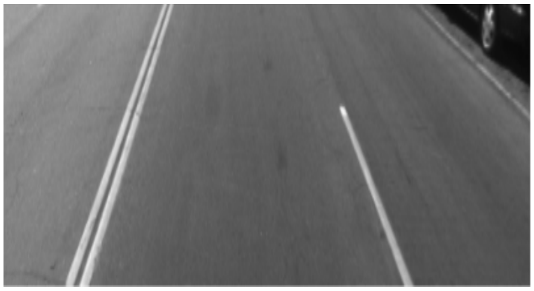


Figure 3. Conversion color image to grayscale image

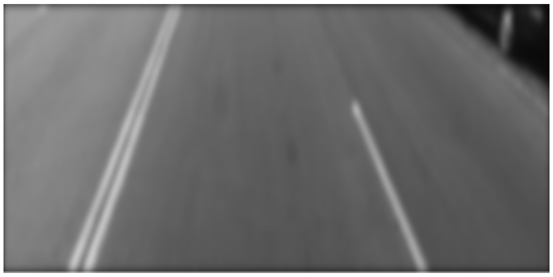


Figure 4. Denoising grayscale image using gaussian filter

3.3. Reducing illumination using Retinex Algorithm

Lighting problem is crucial in Lane Detection System (LDS). In this case, Retinex Algorithm is applied for all types of illumination. The Retinex theory motivated by Land [10] is based on the physical imaging model, in which an image *I*(*x, y*), it could achieve sharpening with compensation for the blurring introduced by image formation process. Moreover, it could improve consistency of output as illumination changes.

In most Retinex methods, the reflectance R is estimated as the ratio of the image I and its smooth version which serves as the estimate of the illumination L.

$$R_i(x,y) = \log I_i(x,y) - \log[F(x,y) * I_i(x,y)] \quad (1)$$

$$R_i(x,y) = \log \frac{I_i(x,y)}{F(x,y)*I_i(x,y)} = \log \frac{I_i(x,y)}{\bar{I}_i(x,y)} \quad (2)$$

where $I_i(x, y)$ is the image distribution in the ith spectral band and R_i (x, y) is retinex output.

Gaussian function F(x, y) =K $e^{-(x2+y2)/c2}$ where K is determined by

$$\iint F(x,y)dxdy = 1 \quad (3)$$

3.4. Detecting Line Segments Angles Computation based on Hough Transform

In this stage, edge information is the most commonly used features for lane boundaries detection system. In real-time situations, lane edge features may not be strong and may be affected by different lighting conditions. Therefore, the selection of edge detection operator is needed. In this case, sobel operator, canny operator, prewitt operator, Roberts, Laplacian of Gaussian and Zero-cross are individually experimented as edge detection algorithms as shown in Fig.9.

From our experiments, sobel operator is the most suitable one for next steps of lane detection system. Thus, this operator is utilized as edge detection algorithm for our proposed system.

Then, Hough Transform is applied for lane detection. The key to generalizing the Hough algorithm to arbitrary shapes is the use of directional information. Given any shape and a fixed reference point on it, instead of a parametric curve, the information provided by the boundary pixels.As Fig.5 illustrates, Hough Transform detects many unnecessary line segments. Therefore, we need to remove the unnecessary line segments for reducing the complexity and computational time. In our research , the angle of a lane boundary is utilized to remove the unnecessary lane segments. After applying angle computation, we obtain left and right boundaries with the correct lines as shown in Fig.7.

For left boundary,

$$15°˚ \leq \theta_{left} \leq 70˚°$$

For right boundary,

$$- 70˚° \leq \theta_{right} \leq -15˚°$$

where θ_{left} is the angular area and θ_{right} is the orientation angle between two lines.

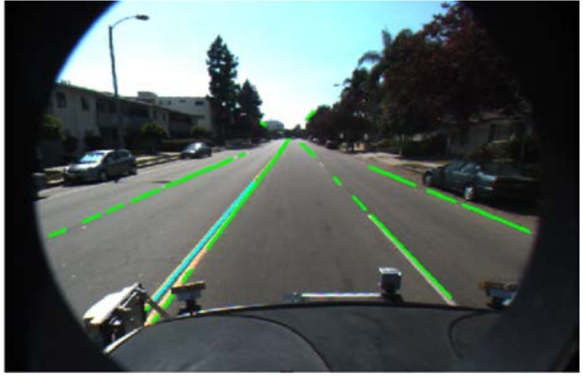


Figure 5. Detecting incorrect lines using only Hough Transform

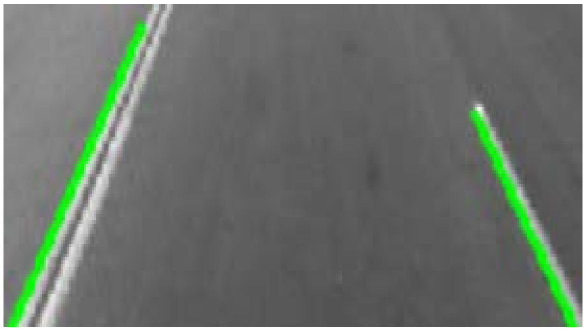


Figure 6. Detecting left and right boundaries in region of interest (ROI)

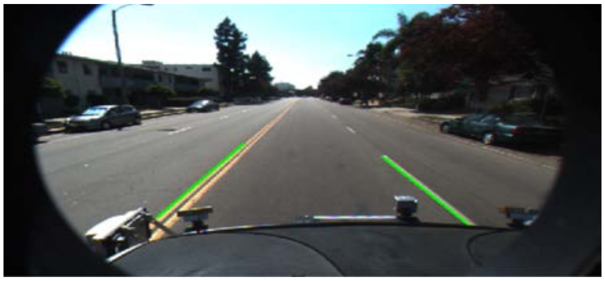


Figure 7. Detecting correct lane boundaries in original image

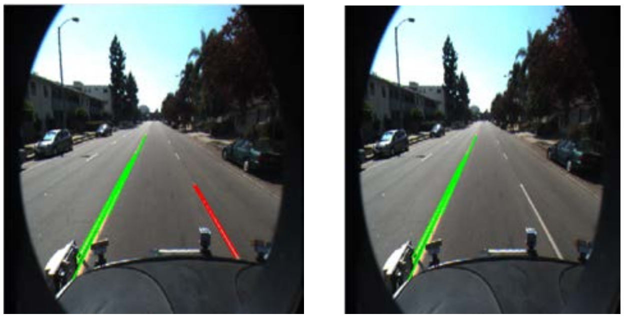
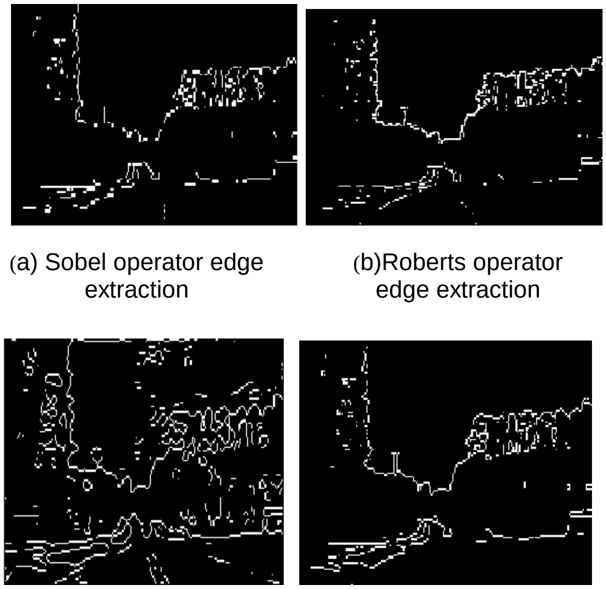


Figure 8. Unclear lane boundaries due to out of bound θ values

4. Experimental Results

We have evaluated the proposed algorithms with a laptop, in Matlab environment which has Intel® Core (TM) i5-7360U CPU@ 2.30 GHz and 8.00GB RAM. Since Caltech's 2008 are available as public database, we have implemented our experiments on it.

In our proposed system, four stages are presented: determining region of interest (ROI), preprocessing, reducing illumination conditions and finding line segments by segments' angles computation based on Hough Transform as shown in Fig.1. The goal for our lane boundaries recognition system is to solve lighting problem for all types of illumination.

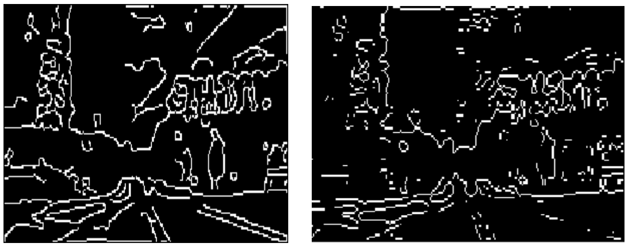


(a) Sobel operator edge extraction

(b)Roberts operator edge extraction

(c) Zero-cross operator edge extraction

(d) Prewitt operator edge extraction



(e)Canny operator edge extraction

(f) Laplacian of gaussian operator edge extraction

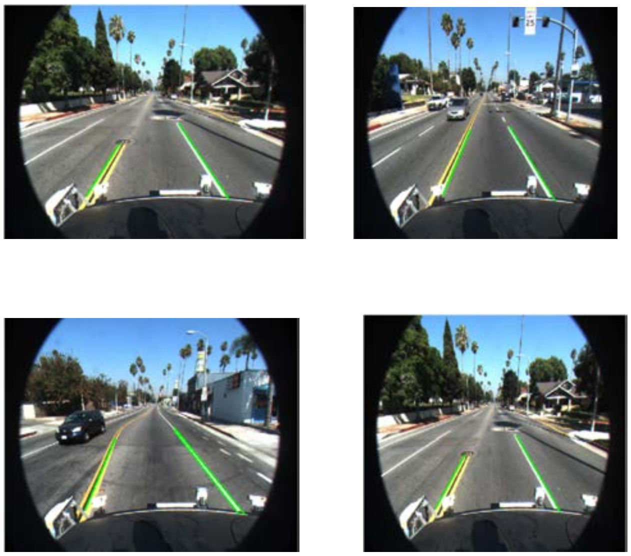
Fig.9.Results of edge extraction methods

The efficient lane boundaries recognition method based upon Retinex with Segments' Angles Computation based on Hough Transform is proposed in our paper. Gray scale conversion and removal of noise are applied in preprocessing stage.

In edge detection, sobel operator is the most suitable edge detector for our proposed system. And some illustrations of accurate lane boundaries recognition are described in Fig.10. Experimental results show recognition rate of our method for different illumination status inside Table II.

Table I. Computation time of edge extraction methods

No	Edge Detection Methods	Time(seconds)
1	Sobel	0.414996
2	Roberts	0.520003
3	Zero-Cross	0.543144
4	Prewitt	0.551883
5	Canny	0.623103
6	Laplacian of Gaussian	0.660884



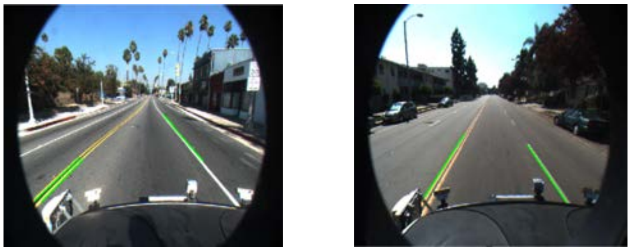


Fig.10. Examples of lane boundaries detection using segments angles computation based on Hough Transform

Table II. Performance statistics of our proposed method

Database	Frame numbers	Accuracy(%)
1	336	85.12
2	231	81.36
3	249	80.12
4	405	86.42

5. Conclusion and Future Work

Lane boundaries detection system faces lighting problem for various types of illuminations in real-time situations. In this paper, we have especially described an efficient lane boundaries detection system based on retinex with line segments' angles computation. Experimental results show that our proposed method can detect correct lane boundaries in various lighting conditions. In future work, we consider curve detection based on our proposed method.

6. References

[1] H.Y.Cheng, B.S.Jeng, P.T.Tseng, and K.C.Fan, "Lane DetectionWith Moving Vehicles in the Traffic Scenes", IEEE Trans Intell.Transp.Syst, vol.7, no.4, 2006, pp.571-582.

[2] F.Mariut, C. Fosalau and D.Petrisor, "Lane Mark Detection Using Hough Transform",International Conference and Exposition on Electrical and Power Engineering, IEEE, 2012,pp.871-875.

[3] A.G.Mohapatra, "Computer vision based smart lane departure warning system for vehicle dynamics control",Sensors & Transdueers Journal,vol.132(9),September 2011,pp.122-135.

[4] G.Liu,S.Li, and W.Liu, "Lane detection algorithm based on local feature extraction", Chinese Automation Congress(CAC),2013,pp.59-64.

[5] Jie Guo, Zhihua Wei, Duoqian Miao, " Lane Detection Method based on Improved RANSAC Algorithm",Twelfth

International Symposium on Autonomous Decentralized Systems,IEEE,2015.

[6] Y.Wang, I.Bai, F.Michael,“Robust road modeling and tracking using condensation”, IEEE Transactions on Intelligent Transportion Systems 9 (2008) 570-579.

[7] Hao Yu, Yule Yuan, Yueting Guo, Yong Zhao, “Vision-based Lane Marking Detection and Moving Vehicle Detection”, 8th International Conference on Intelligent Human-Machine Systems and Cybernetics,IEEE,2016.

[8] A.Guiducci, “Parametric model of the perspective projection of a road with applications to lane keeping and 3d road reconstruction”, Computer Vision and Image Understanding 73,414-427, 1999.

[9] Z.Teng, J.H.Kin and D.J.Kang, “Real-time Lane detection by using multiple cues”, IEEE International Conference on Control Automation and Systems, 2010, pp.2334-2337.

[10] M.Aly, “Real time detection of lane markers in urban streets”,Intelligent Vehicles Symposium,IEEE,2008,pp.7-12,4-6.

Internet of Things

The Home Network Sentinel System

Zaw Myint Naing Oo, Khin Kyawt Kyawt Khaing
University of Information Technology
zawmyintnaingoo@uit.edu.mm, khinkkkhaing@uit.edu.mm

Abstract

In recent year, artificial intelligence technology has developed rapidly, and deep learning has been widely used in many areas. This paper presents a home network sentinel system by using deep neural network and this system assured the safety of home appliances that can be integrated with the existing home automation. This system consists of generic rules, a set of rules and inferences the home safety services according to the sensor input values. This system has some filters, that judge is the appropriateness of sensor values. This system will usethe competing home appliances which are automatically controls near boundary on/off. So, this system has the special database system for competing appliances.The Geo Fencing system will also be applied to watch the movement of object. This database system will link to the deep neural network. The Remote UI will be applied to monitor the condition of the home status.

Keywords-Home safety, Generic Rules, Database, Geo Fence, Deep Neural Network, The Remote UI

1. Introduction

Intelligent home is an integrated system in a home that integrates multiple home services, where the technology and process used to create a building that can act intelligently so that a home becomes safer and more productive for users and more efficient for its owners.The proposed system realizes by the sensors value rules. This system needs to define the set of fuzzy rules.People are always worried about what would be the condition of their homes and offices when they are not there. Therefore, this proposed system is trying to make a system which would automatically provide the user to save and control the home appliances [1].

There are many types of safety problems that may arise within a home environment. These safety problems can be classified into three big categories: safety of home appliances, safety of indoor environment and safety of interaction between home users and home appliances. The occurrence of home safety problem always have three bad consequences: cause casualty or cause home property loss or both [2].

This system will define the generic rules based on the sensor values. The generic rule is the representation of common senses in terms using the syntax of the decision support system. The system will use some filters to judgment between sensors and embedded appliances, because they might be malfunctioned [3].

Thissystem will use the deep neural network to evaluate the home safety system. Intelligent home is an integrated system in a home that integrates multiple home services, where the technology and process used to create a building that can act intelligently so that a home becomes safer and more productive for users and more efficient for its owners.

The Geo Fencing system will watch and act as a sentinel system, how the user is moving which way the user moves to the Geo Fence (from inside, outside, inside/outside cross direction). These three conditions will include for the Geo Fencing system.

The Remote UI will also provide the functions to monitor and to control the status of home appliances.

2. Related Works

The architectureof rule verification system (RVS) will be taken into rule verification in creating module and evaluated its usability. The rule is extracted and transformed to verification-oriented formal rule representation. The content anomaly detection takes action on the transformed rule and finds whether the content of the rule is inconsistence with the domain knowledge in knowledge base. Creating Conflict Detection(CCD) compares the rule with other related and analysis. If conflicts exist, the CCD works out the conflict degree and generate report to user at the same time, if not exist, the new rule will be added to RVB. On the other hand, after the rule gets triggered and prepares to execute, it will be sent to Rule Verification in Executing (RV_E) module. Executing Rule Verification (ERV) detects rule conflict again to make sure the execution of this rule will not bring the system disorder. Conflict Resolution (CR) modules works as the assistance of the two rule verification modules RV_C and RV_E. CR provides rational conflict resolution strategies in case of rule conflict to improve the system effectiveness [4].

Intelligent home management system has been developed which has the ability to turn on and turn off the room lights automatically, record the controlled electronic devices usage status, switching on and off air condition regulating device automatically, showing temperature room in the house, detect fire signs in the house and turned on the sprinklers in the home in case of fire, supervising the home through surveillance cameras, storing photos and surveillance records on home, detecting people movement in home, and providing notification when someone entered home. System is implemented in prototype. The results show that the system can detect light intensity, flame, room temperature, movement of people, and home state and then the information is successfully sent to the server over the WiFi. The result can be read from server by using browser and there is a data logger in the server. Intelligent home management system prototype development covers hardware and software implementations [5].

Expert systems are normally used in various problem solving and decision making activities such as monitoring, diagnosing and various training related activities. Yashwant Singh Patel proposed a framework that is based on wireless sensors and expert system to solve day to day problem occurring in home appliances. Whenever problem occurs in any part of home appliance, the sensor detects that problem automatically and sends it for solution to the expert system, various noise removal algorithms for removing noise from the received data can be applied for getting noise free data. The expert system finds the solution based on the type of problem and sends the solutions with various images through SMS or e-mail to user's mobile or mail-id [6].

This monitoring system proposed a home automation environment with a controller device that receives data and control information from connected sensors and actuators. The controller devices are attached to a monitoring device, which receives information of device behaviour based on events. The monitoring device includes a rule execution engine evaluating rules in real time and storing their state (satisfied, temporarily violated and permanently violated), according to events received from the controller device.

The consistency checker evaluates the consistency of these rules and reports if their execution is compatible. The action performer executes some actions triggered by the executing rules, depending on their current state. The architecture also includes a mechanisms for rule composition and publication in the form of services over a cloud infrastructure. The rule execution engine can download these services and execute them in the monitoring device [7].

3. Background Theory

3.1Deep Neural Network

A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modeling to process data in complex ways [8]. Deep Neural Networks (DNN) is an important method for machine learning and has been widely used in many fields[12].

3.2. Geo fencing

Geo fencing is a technology used to monitor mobile objects (vehicles, persons, container, etc.,), located by GPS. The geographic coordinates of the tracked object are automatically and regularly sent to a control center, via mobile phone networks. The set of geographic coordinates is used to constitute a virtual boundary (Geo Fence) around a geographic area. The system can determine whether the tracked object is located inside or outside the Geo Fenced area. This technology can also allow the detection of spatial proximity between the tracked mobiles and a specific Geo Fenced area [9].

3.3 IFTTT

IFTTT is a web based service that allows Internet users to create a chain-reaction from one web service application to another. Based on a user-defined conditional statement, called a recipe, the trigger of one web service application activates an action of another web service application. The IFTTT model can be applied to home automation devices where one device can trigger the action of another device. The IFTTT technology is described as shown in Figure 1. The Figure 1 describes how home automation devices would react on the user-define recipes. Two recipes are shown in Figure 3. First recipe is "If motion is detected in a room, then turn on the lights". When the motion sensor in the room detects a movement, it sends a trigger to the central node. Based on the recipe and the trigger, the central node sends an action to the room lights to turn on. Second recipe is "If temperature and humidity changes in the garden, the turn on the irrigation system". When the temperature and humidity sensor senses change, it sends the trigger to the central node. Then, the trigger is interpreted by the central node that sends an action to the irrigation system. These recipes can be generated by remotely accessing the central node of the home automation system, or it can also be accessed within the home network. The central node acts as a router for the home devices to access the Internet and integrates all different types of data communication mediums. Therefore the central node offers a web interface to allow users to configure the different recipes, which can be accessed from computers, smartphones or tables [10].

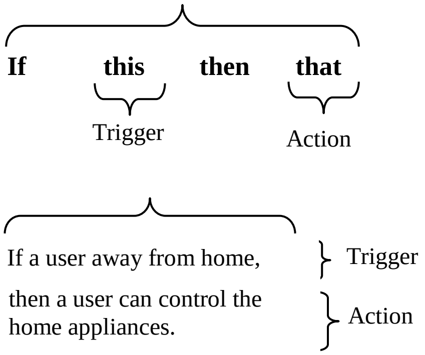

If
this
then
that
Trigger
Action
If a user away from home,
Trigger
then a user can control the home appliances.
Action

Figure 1. IFTTT Description

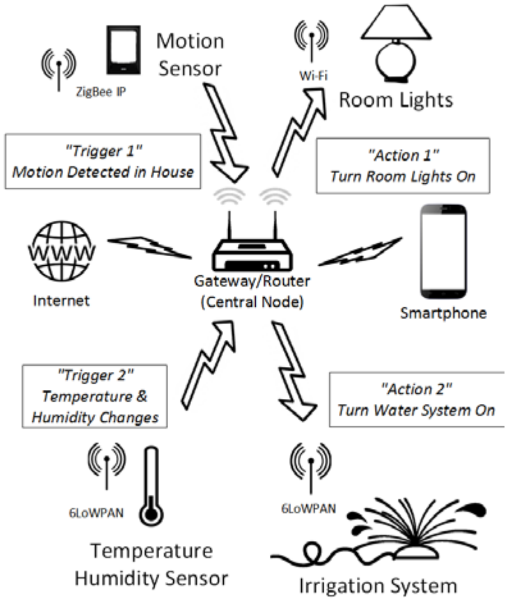

ZigBee IP
Motion Sensor
Wi-Fi
Room Lights
"Trigger 1" Motion Detected in House
"Action 1" Turn Room Lights On
Internet
Gateway/Router (Central Node)
Smartphone
"Trigger 2" Temperature & Humidity Changes
"Action 2" Turn Water System On
6LoWPAN
6LoWPAN
Temperature Humidity Sensor
Irrigation System

Figure 2. Home Automation Overview

3.4 Remote UI

Remote UI refers to Web 2.0. The user can create new services by combining the object provided services, it is called Web 2.0 or mashup. It can be specialized for the composition of services that enable accessing/controlling smart things [10]. A mashup is a web application or a web page which usually uses application programming interfaces (APIs) in order to blend information from multiple sources to create compelling services. As more and more embedded devices (like smartphones and sensor equipped appliances) will be apply to provide their functions as services online, and an abundance of real objects will essentially become a part of ambient spaces (interoperating and communicating over TCP/IP networks), the need to create value-added services by composing numerous embedded-device enable services [11].

4. The Architecture of Home Network Sentinal System

The architecture of home network sentinal system is shown in Figure 3. This architecture includes the special database system for the competing home appliances, some filters to judgement for sensors and home appliances to avoid malfunction, the Geo Fencing rules for intelligence fence, IFTTT acts like as remote control, sensor value rules for controlling the home appliances and the remote UI for monitoring the status of home.

This system will use generic rules. Firstly, the system needs to define a set of rules according to the input sensor values. This system also needs to define competing functions for competing appliances. The input value can be obtained by the sensors. Secondly, it proceeds to deploy the deep neural network. Finally the system will save the home appliances according to the competing functions which located in the special database.

In this system, the size of Geo Fence size can range from a few tens of meters to several kilometers. The Geo Fencing areas can be defined by geometric shapes. The geographical areas are defined as circular area, rectangular area and ellipsoidal area.

This system defines the circular geographical area with a single point that represents the center of the circle and a radius. Coordinates from characteristic points of the shape are necessary to define the Geo Fence perimeter.These coordinates are used in equation (1), along with the inside or outside of the Geo Fence, which enables the computing of alerts.Sensor value rule uses the appropriate sensor values within the total range and the geo-fencing rules use fuzzy control logic, which is the IF THEN statements. The geographical circular area is described as shown in Figure 4.The function of geographical circular area is defined by equation (1).

$$F(x, y) = 1 - \left(\frac{x}{r}\right)^2 - \left(\frac{y}{r}\right)^2 \quad (1)$$

Where F is the function to determine the spatial characteristics of a point (x,y) relative to a geometric shape, r is the radius of a circle, x is the abscissa of a Cartesian coordination system with the origin in the center of the geographical area, y is the ordinate of a Cartesian coordination system with the origin in the center of the geographical area. The function F defined in equation (1), determines whether a point is located inside, outside, at the center, or at the border of a geographical area.

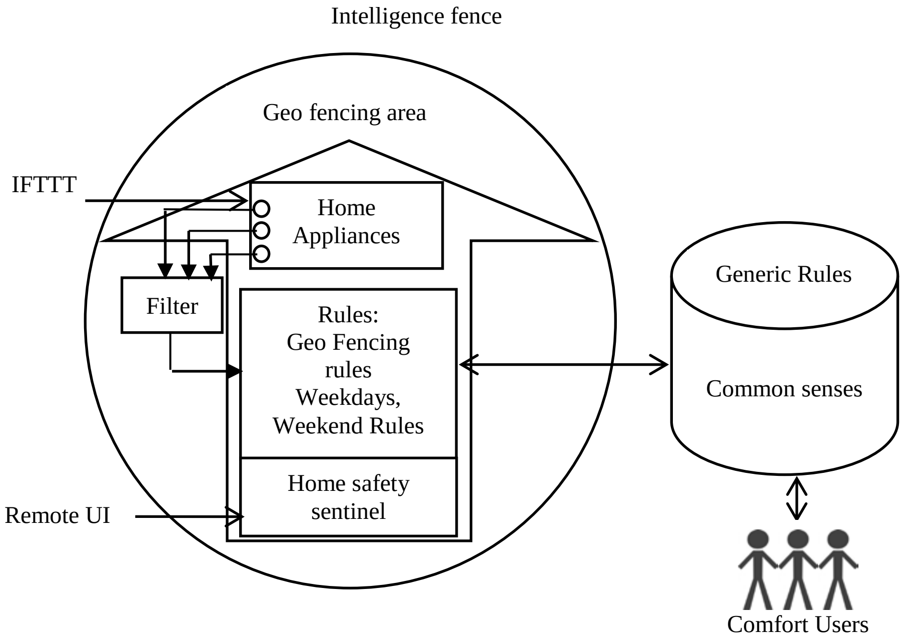


Figure3. The architecture of home network sentinel system

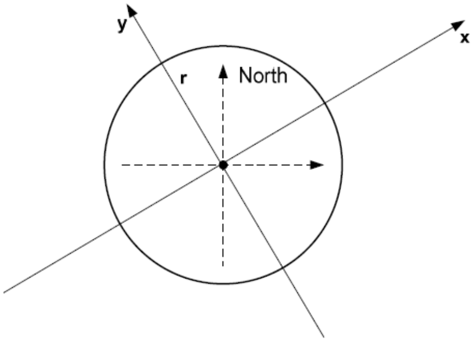


Figure 4. The geographical circular area

If the value of function F is equal to one, the location is at the center point of the geographical area. If the value of function F is greater than zero, the location is inside the geographical area.If the value of function F is equal to zero, the location is at the border of the geographical area.If the value of function F is less than zero, the location is outside the geographical area.

IFTTT is a web-based service that allows Internet users to create a chain-reaction from one web service application to another. Based on the IFTTT (IF-This-Then-That) model, this system will define a set of device communication protocols where devices' triggers and actions are combined to manage interactions for home safety. This system uses Web 2.0 for remote user interface and creates new services by combining the object provided services.

5. Evaluation of Home Safety System

The sensor and embedded appliances might be malfunctioned. So, this system will define the generic rulesand must have some filters to avoid the malfunctions.The generic rule is the representative of common sense. The scenario is using the temperature to define the rules. These rules will be putted into the special database. When the people are sleeping sometime, they use the blanket, because the room temperature is a little bit low, i.e., common sense. There has rules based on the common senses, about the temperature, humidity, the electricity usages and more interestingly the home appliances, we can hold the state of appliances then we will be doing some more interesting senses, competing appliances (e.g., the heater and air con).

In this system, we assume that the competing appliances, the air conditioner is set to 25 degree and the heater is also set to 25 degree. When the heater starts to heat, it takes time to give warm. When the room temperature is high, the air conditioner kick the heater, it takes time to low the room temperature. They may over shift belong, and as soon as over shift. But, the temperature was higher than 25 degree, may be outside temperature as like 30 degree, at that time the air conditioner kick the heater, to cool down the temperature, (i.e competing appliances (competingfunctions)).

The following rules are the generic rules to use the competing functions. These rules are located into the special database.

Rule 1: if the heater is higher than 25 degree, then the air conditioner is cool down the temperature until 25 degree.

Rule 2: if the air conditioner is lower than 25 degree, then the heater is high temperature until 25 degree.

Rule 3: if the outside temperature is higher than 25 degree, then the air conditioner is cool down the temperature until 25 degree and the heater is just warm.

Rule 4: if the outside temperature is lower than 25 degree, then the heater is warm up to 25 degree and the air conditioner is still 25 degree.

This system will also use the rules for the Geo Fencing system. In this system, it uses set of rulesfor testing the Geo Fencing system which is shown in table 1.

Table 1. Rules for testing the Geo Fencing System

	Light	Air Con	Fan	Doors
Inside	On	On	On	On
Center	On	On	On	On
Border	On	On	On	On
Outside	Off	Off	Off	Off

This system defines the following rules to control the home appliances by using the geographical area.

Rule 1: If F(x,y) = 1 then the location is at the center point of the geographical area (control the home appliances)

Rule 2: If F(x,y) > 0 then the location is inside of the geographical area (control the home appliances)

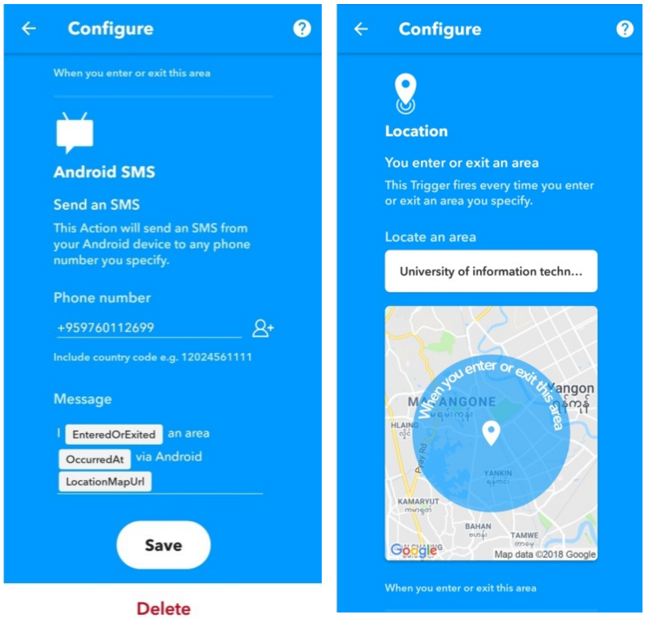
Rule 3: If F(x,y) = 0 then the location is at the border of the geographical area (control the home appliances)

Rule 4: If F(x,y) < 0 then the location is outside of the geographical area (lock the home)

The Figure 5 is showing the IFTTT service, how to configure and how to control the appliances by using IFTTT service. The IFTTT service can create chains of conditional statements, which is called ‘applet’. The following conditional statement is tested based on android location, which acts as a remote control for appliances.

If (EnteredOrExited) an area (OccuredAT) via Android (LocationMapUrl) then (Notify or Control the appliances)

If the user entered or exited at the specified area then send notification to the user and the user can control appliances which the user wants to switch on/off for electrical appliances.



Configure

When you enter or exit this area

Android SMS

Send an SMS

This Action will send an SMS from your Android device to any phone number you specify.

Phone number

+959760112699

Include country code e.g. 12024561111

Message

I EnteredOrExited an area OccurredAt via Android LocationMapUrl

Configure

Location

You enter or exit an area

This Trigger fires every time you enter or exit an area you specify.

Locate an area

University of information techn...

When you enter or exit this area

Figure 5. IFTTT service

6. Conclusion

This paper proposes a home network sentinel system which helps us to assure the safety of home appliances and home environment. This system acts as a sentinel, which knows the movement of user to Geo Fence from inside, outside or cross direction. It can provide home safety functions.In this system has a special database for competing appliances. It is a new technique of implementing home safety system that will give more safety for smart home appliances based on the rules. This system will save cause casualty or cause home property loss or both. There exists several home safety systems. This system to be more effectively and safety for home. This system will be acted intelligently the home safety services as like the human manner.In future work, this research plans to develop the more detail of sentinel system for home appliances.

7. References

[1] Zaw Myint Naing Oo, Tha Pyay Win, “The Development of an Intelligent Fuzzy Expert System for The Home Safety System”, The 15th International Conference on Computer Application2017, Feb 16th -17th 2017, 13-18.

[2] Zhengguo YANG, Azman Osman LIM, Yasuo TAN, School of Information Science, Japan Advanced Institute of Science and Technology, “Event-based Home Safety Problem Detection Under The CPS Home Safety Architecture,” 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)

[3] Zaw Myint Naing Oo, Khin Kyawt Kyawt Khaing, “ The Home Safety System Based on Competition Functions”, The 1st International Conference on Advanced Technologies ICAIT 2017, 139-144

[4] Hong Luo/Ruosi Wang, Xinming Li, “A Rule Verification and Resolution Framework in Smart Building System”, 2013 International Conference on Parallel and Distributed System, 1521-9097/13 2013 IEEE, DOI 10.1109/ICPADS.2013.74

[5] Azka Ihsan Nurrahman, Kusprasapta Mutijarsa, “Intelligent Home Management System Prototype Design and Development,” International Conference on Information Technology Systems and Innovation, Bandung-Bali, November 16-19, 2015 IEEE

[6] Yashwant Singh Patel, Sneh Vyas, Atul Kumar Dwivedi, “A Expert System based Novel Framework to Detect and solve the Problems in Home Appliances by Using Wireless Sensors,” 2015 1st International conference on futuristic trend in computational analysis and knowledge management (ABLAZE 2015)

[7] Ramon Alcarria, Diego Martin de Andres, Borja Bordel, Diego Sanchez de Rivera, Alvaro SanchezPicot, Tomas Robles, “ A Service-Oriented Mornitoring System Based on Rule Evaluation for Home Automation”, 2017 IEEE International Conference on Consumer Electronic (ICCE), 978-1-5090-5544-9/ 2017 IEEE

[8] https://www.techopedia.com/definition/32902/deep-neural-network

[9] Fabrice RECLUS, Kristen DROUARD, “Geofencing for Fleet and Freight Management,” 2009 IEEE

[10] Thomas Gonnot, Won-Jae Yi, Ehsan Monsef, Jafar Saniie, “Home Automation Device Protocol[HADP]:A Protocol Standard for Unified Device,” Advances in Internet of Things, 2015, 5, 27-38

[11] Nikos Vesyropoulos and Christos K. Georgiadis, “ Custmized QoS-based Mashups for the Web of Things: An Application of AHP,” Computer science and information systems 12(1):115-13

[12] Huang Yi, Duan Xiusheg, Sun Shiyu, Chen Zhigang, “A Study on Deep Neural Networks Framework,” 978-1-4673-9613-4/16 ©2016 IEEE

Natural Language Processing

Finding Myanmar Word Similarity by Word Embedding

Myat Sapal Phyu
University of Information Technology, Yangon, Myanmar
myatsapalphyu@uit.edu.mm

Khin Thandar Nwet
University of Information Technology, Yangon, Myanmar
khinthandarnwet@uit.edu.mm

Nwe Ni Aung
University of Information Technology, Yangon, Myanmar
nweniaung@uit.edu.mm

Abstract

Word embedding is very efficient vectorizer that converts word to numerical vector by capturing the semantic relation with context words. It is one of the most dominant development in Natural Language Processing (NLP). Although word embedding model can be applied without anxiety for data collection and preprocessing in English language, well preprocessing is needed for Myanmar language. Text preprocessing is crucial to the construction of word embedding model and it is significantly effect on final result. We prepare 252,842 words from three different Myanmar news topics (health, crime and education) by keeping 2-gram to 7-gram. This paper investigates the analysis of detecting the similarity or relatedness between Myanmar words by skip-gram model.

Keywords - Word embedding, Myanmar news, Natural Language Processing (NLP), skip-gram.

1. Introduction

Word embedding model is an active research area and its applications have recently great attraction in the trend of Natural Language Processing (NLP). Word embedding is the conversion of word into numerical representation of contextual similarities between words by capturing the meaning and semantic relationships. Generally, word embedding can be classified into two groups, frequency based embedding and prediction based embedding. One of the most popular method to construct word embedding model was proposed by Mikolov et al., [5, 6] implemented in Word2Vec tool. Word2Vec is the combination of two techniques, Continuous Bag of Words (CBOW) and skip-gram model. CBOW predicts the probability of a word by given context. Skip-gram predicts the context of word by given word. Word2Vec model is a prediction based embedding. Word2Vec remains popular due to their efficiency and simplicity. The purpose of Word2vec is to group the similar words vector together in vector space. Word2Vec can guess highly accurate about a word's meaning with enough data. Technically, Word2Vec can not be considered as the part of deep learning because the number of parameters and layers are too small to be considered a deep learning model. It can transform text to numerical form that can be understand by deep neural network. The output of the Word2Vec is a vocabulary with corresponding vector for each word that can expect the relationship between words. The rest of the paper is as follows, section 2 discusses the related researches that was published in the area of word embedding. Section 3 describes the overall system architecture, section 4 describes data collection and preparation. Section 5 explains the preprocessing task including syllable segmentation, word extraction, removing stop words. Section 6 explains about word embedding especially focus on skip-gram model. Section 7 describes the experimental result and Section 8 concludes the paper.

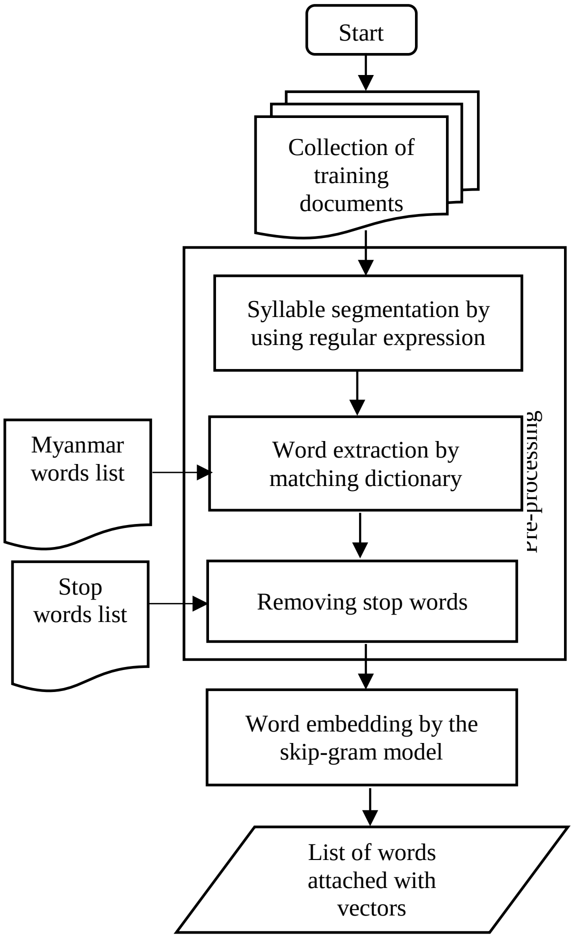
2. Related Works

In this section, we investigate the long history of word embedding. Since the 1990s, continuous representation of words in vector space were estimated by many models including Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). LDA predicts a word from a global context. LSA is based on several parameters, local and global frequencies, local and global weighting functions and the dimension of the semantic space. Naili et al. [7] proved that the quality of topic segmentation depends on these parameters. They identified the best combination of LSA parameters : local frequency = 3, global frequency = 1, local weighting = TF , global weighting = IDF and dimension of the semantic space = 70%. Density function term frequency–inverse document frequency (tf-idf) is a weighting factor that is used to detect important words in the collection of documents or corpus. Tf-idf is also a vectorizer that convert terms or words to numerical vector. Many variations of tf-idf weighing scheme [1, 4] are often used in search engine, text summarization and classification, etc. These techniques can be considered as the most influential early models for word embedding.

In recent years, Word2Vec model by Mikolov et al., [5, 6] become the most popular and efficient word embedding model and many researchers investigate and experiment with Word2Vec and similar techniques [8, 10]. The advantages of word2vec over early models is that it can convert high dimensional vector into low dimensional vector and it can maintain word context. Word2Vec contains two model architecture Continuous Bag of Word (CBOW) and skip-gram model. In this paper, skip-gram model is used because of its simplicity and effectiveness.

3. System Architecture

Figure 1. shows the overview of the system. Firstly, Myanmar text data are collected from Myanmar daily news websites [11, 12]. Since the purpose is to convert word to vector, it is needed to extract word from the collection of text documents. As the preprocessing step for word extraction, syllable are segemented from the input text document. After that, segmented syllables are merged by matching dictionary. Then, unnecessary words are removed by matching stop words list. Finally, extracted words are fed into skip-gram model in order to find the similar contexts of each word by converting numerical vectors.



Start
Collection of training documents
Syllable segmentation by using regular expression
Myanmar words list
Word extraction by matching dictionary
Pre-processing
Stop words list
Removing stop words
Word embedding by the skip-gram model
List of words attached with vectors

Figure 1. Overview of the system

4. Data collection

In English language, text data can be used from many standard datasets. In Myanmar language, text data are collected from Myanmar daily news website [11, 12] and extracted text by online text extractor [13]. Then, extracted text are converted into unicode format by zawgyi to unicode converter [17]. Each news article is saved as text document (.txt) and used as input data set.

Each news article generally contains about 10 sentences. In this paper, 683 news articles for crime category, 620 news articles for education category and 683 news articles for health category are collected as text documents. Table 1. shows the data set on three different news categories. Words are extracted from collection of text after some perprocessing steps, 16,381 features from 683 crime news, 14,801 features from 620 education news and 12,273 features are extracted from 683 health news. Total number of sequence of words 252,842 will be used as input data for the construction of word embedding model.

Table 1. Data set on three online news

No.	Category	Number of Documents	Number of Extracted Features
1.	Crime	683	16,381
2.	Education	620	14,801
3.	Health	683	12,273
Total			252,842

5. Preprocessing

Large amount of words is needed in order to construct word embedding model. Some necessary preprocessing tasks for English text data can be performed easily by using and installing existing tools and libraries. Extraction of words from English text corpus can be performed easily by detecting word boundary with white space. In order to extract words from the collection of Myanmar text, it is firstly needed to segment text into separate syllable and the segmented syllable are merged in order to form a meaningful words.

In this paper, syllables are segmented by syllable segmentation method that is implemented by regular expression pattern [15]. The pattern is based on encoding order of Myanmar syllable. Segmented syllables are merged by matching dictionary [16]. In this paper, 2-gram to 7-gram is used to keep all compound words and simple words that constitute in the compound word such as "ဖမ်းဆီးရမိ, ဖမ်းဆီး, ရမိ, ကျန်းမာ, ကျန်းမာရေး". It can be more convenient to find the semantic relations of words and also the alternative of increasing the size of training data like some regularization strategies that are used in training state. Figure 1. shows the sample of preprocessing task.

After extracting word from text documents, unnecessary words are remove by matching stop word list. In this work, stop words are collected by analyzing Myanmar online news. Most of the news contains the location and time information that are not important terms for categorizing news documents. After analyzing news documents, location, date, time words and the most commonly used prepositions, inflections; conjunctions are collected as stop words. Moreover, punctuation marks (eg., "။၊ ."), white spaces and other symbols (eg., "-/()[]{}"), non-Myanmar text (eg; A to Z), numerical text (eg., 0 to 9 ၀to ၉) are removed. In this paper, 608 stop words are collected and more stop words will be added in the future. Table 1. shows the sample of stop words list.

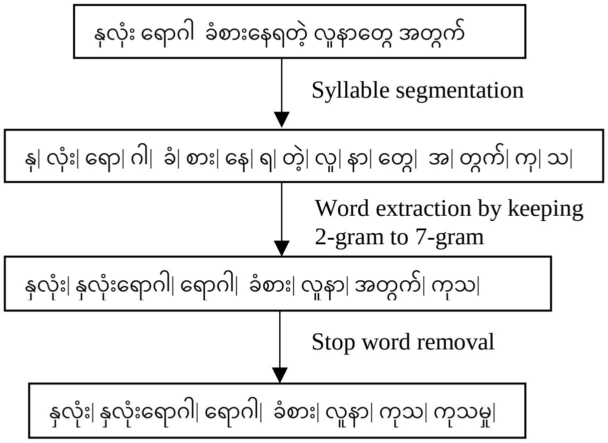

နှလုံး ရောဂါ ခံစားနေရတဲ့ လူနာတွေ အတွက်
Syllable segmentation
နှ| လုံး| ရော| ဂါ| ခံ| စား| နေ| ရ| တဲ့| လူ| နာ| တွေ| အ| တွက်| ကု| သ|
Word extraction by keeping 2-gram to 7-gram
နှလုံး| နှလုံးရောဂါ| ရောဂါ| ခံစား| လူနာ| အတွက်| ကုသ|
Stop word removal
နှလုံး| နှလုံးရောဂါ| ရောဂါ| ခံစား| လူနာ| ကုသ| ကုသမှု|

Figure 2. Sample of preprocessing task

Table 2. Sample of stop words

Date, time	ပြေ၊ ဇွန်၊ ဇူလိုင်၊ ဩဂုတ်၊ နာရီ၊ မိနစ်၊ စက္ကန်၊ နေ့လည်၊ နတ်တော်၊ ပြာသို၊ တပို့တွဲ
Location	တိုင်းဒေသကြီး၊ မြို့နယ်၊ မြို့သစ်၊ ကျေးရွာအုပ်စု၊ ဘိုကလေး၊ ဓနုဖြူ၊ ဒေးဒရဲ၊ မင်္ဂလာဒုံ။
Conjunction	သည့်အပြင်၊ ထို့ပြင်၊ ထို့အပြင်၊ ဒါ့အပြင်၊ နောက်ပြီး။

6. Word embedding

Word embedding converts the words into vectors in low dimentional space. There are different kinds of word embedding including count vector, tf-idf vectorization, word2vec that contain two model architecture, Continuous Bag of Words (CBOW) and skip-gram model and GloVe. Anyhow, the main target of word embedding models is to transform text into number in order to reduce dimension and maintain contextual similarity. In this paper, we focus on skip-gram model because of its simplicity and effectiveness. Word embeddings can be manipulated the degree of similarity between two words, for instance, similarity('boy', 'girl') as 0.7452678. Such kind of similarity is measured by word similarity. They can be also manipulated thing like the query is most_similar(positive=['woman', 'king'], negative=['man'], topn=1) and the result is queen. The accuracy can be measured by word analogy. Word analogy accuracy can be computed based on question-answer pair in the form of a:b::c:?. In this paper, evaluation is performed only on word similarity because of rare of language resources to prepare and reference for question-answer pairs.

6.1. Skip-gram model

It trains on simple neural network with a single hidden layer to predict context words from a word. The training objective of skip-gram model is to learn word vector representation that predict nearby words. It predicts nearby context words from the continuous sequence of training words w_1, w_2, w_3, …….., w_T. The objective of skip-gram model is for maximizing the average log probability. [6]

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c,j\ne 0}\log p(w_{t+j}|w_t) \tag{1}$$

The basic skip-gram formulation using softmax function:

$$p(w_O|w_I)=\frac{\exp({v'_{w_O}}^{\top} v_{w_I})}{\sum_{w=1}^{W}\exp({v'_{w_o}}^{\top} v_{w_I})} \tag{2}$$

Where,
v'_{w_o}= output vector representation of w
v_{w_I} = input vector representation of w
W= number of words in the vocabulary

Let consider the following sample, assume that we use contex window size 2 and the target word is "ခံစား" then (ရောဂါ, ခံစား), (နှလုံး, ခံစား), (လူနာ, ခံစား), (ကုသ, ခံစား) of (context, target) pair. Figure 3. depicts sample of continuous word sequence data with context window size 2.

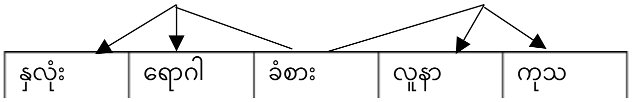


Figure 3. Sample of context size 2 for word "ခံစား"

Architecture of skip-gram model is depicted in Figure 3. Input layer gives a single word X_k=1 and $X_{k'}$=0 for k ≠

k' (one-hot encoding). Hidden layer is represented by V×N matrix W. Each row of W is the N dimensional vector representation of related word w in input layer. The output is calculated by softmax function.

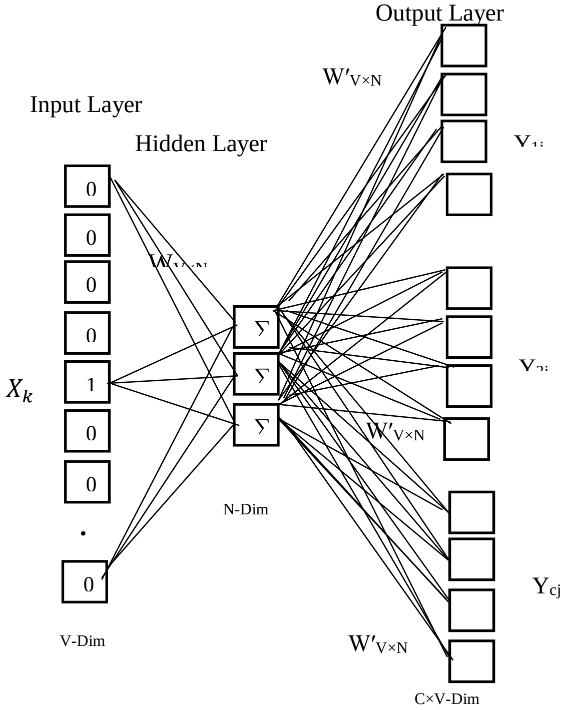


Figure 4. Architecture of skip-gram model

7. Experimental result

In this paper, we use the dataset consisting of various Myanmar news articles including crime, education, and health topics with total 252,842 extracted words. We tested with different vector dimension 50, 100, 200, 300 and context window size 3. Table 3, 4 and 5 list the top five words of "ဒဏ်ရာ-injury ", "လူနာ-patient" and "စာမေးပွဲ-exam" according to the experiment. The similarity score 1 means that two vectors are equal (eg., similarity between "ဒဏ်ရာ" and "ဒဏ်ရာ" is 1) , and 0 means they are bearing no relation to each other. According to the experiment the most similar word of the word "ဒဏ်ရာ" is "ပေါက်ပြဲ" with similarity score 0.7238627076, the most similar word of the word "လူနာ" is "ဆေးရုံ" with similarity score 0.7548751831 but it can be said that related words rather than similar. Then, the most similar word of the word "စာမေးပွဲ" is "တက္ကသိုလ်ဝင်စာမေးပွဲ" with similarity score 0.7208846986.

Table 3. Top five words similar to "ဒဏ်ရာ"

Word	Similarity [0-1]
ပေါက်ပြဲ	0.7238627076
ထိခိုက်	0.7160881757
ပြတ်ရှဒဏ်ရာ	0.7128141522
ဒဏ်ရာရ	0.7058121562
ထိုးသွင်းဒဏ်ရာ	0.6930733919

Table 4. Top five words similar to "လူနာ"

Word	Similarity [0-1]
ဆေးရုံ	0.7548751831
သေဆုံး	0.7282525300
ရောဂါ	0.6954033195
သွေးလွန်တုပ်ကွေး	0.6889818310
တုပ်ကွေးရောဂါ	0.6605198383

Table 5. Top five words similar to "စာမေးပွဲ"

Word	Similarity [0-1]
တက္ကသိုလ်ဝင်စာမေးပွဲ	0.7208846986
အတန်းတင်စာမေးပွဲ	0.7108163309
အောင်စာရင်း	0.6986297928
ကျရှုံး	0.5730306267
ဘာသာစုံဂုဏ်ထူး	0.5509787976

In this paper, the performance of the skip-gram model is measured by intrinsic word similarity. In word similarity evaluation, a list of two pairs of words with their similarity rating judged by human annotators is used to calculate the accuracy of the model. In English language, word similarity is evaluated by matching word similarity dataset including wordsim-353 [2], MTurk-287 [9], SimLex-999 [3] and so on. So, we collect 330 words with 2884 similar word pairs in Myanmar language by analysing the relatedness and similarity of word and by referencing Myanmar synonym book [18] to calculate the accuracy of the skip-gram model. Most of the Myanmar words have many synonym words, different words with similar meaning. Each word has at leat five synonyms. Table 6. shows the sample of synonyms list for one word "ကြိုးစားပမ်းစား", means "great effort" in English.

Table 6. Synonyms of “ကြိုးစားပမ်းစား”

Word 1	Word 2
ကြိုးစားပမ်းစား	ကြိုးကြိုးစားစား
ကြိုးစားပမ်းစား	ကြိုးကြိုးပမ်းပမ်း
ကြိုးစားပမ်းစား	အပတ်တကုတ်
ကြိုးစားပမ်းစား	ကြိုးကြိုးကုတ်ကုတ်
ကြိုးစားပမ်းစား	အားကြိုးမာန်တက်
ကြိုးစားပမ်းစား	အားသွန်ခွန်စိုက်
ကြိုးစားပမ်းစား	သဲကြီးမဲကြီး
ကြိုးစားပမ်းစား	သဲသဲမဲမဲ
ကြိုးစားပမ်းစား	သစ်ခုတ်ကျားစီး
ကြိုးစားပမ်းစား	ကျားကုတ်ကျားခဲ

Table 7. shows the performance evaluation of skip-gram model by word similarity for three domains including crime, health and education news.

Table 7. Evaluation measure by skip-gram model

	Precision	Recall	F-Score	Accuracy	Error Rate
Crime	89	87	88	75	18
Health	81	84	82	63	22
Education	79	81	79	60	24

8. Conclusion

This paper especially focus on the investigation and analysis of skip-gram model for embedding Myanmar words. The objective is to convert Myanmar words into numerical form that deep neural network can understand. It detects the similarities of words to group the vector of similar words together in vector space. It can be made highly accurate guesses about word’s meaning if we use large data usage and context. This will lead to convenient application of Myanmar language for many text classification tasks including sentiment analysis, article classification, spam detection and so on.

9. References

[1] Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016, March). “A novel text mining approach based on TF-IDF and Support Vector Machine for news classification”. *In Engineering and Technology (ICETECH), 2016 IEEE International Conference on* (pp. 112-116). *IEEE.*

[2] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001, April). “Placing search in context: The concept revisited”. *In Proceedings of the 10th international conference on World Wide Web* (pp. 406-414). ACM.

[3] Hill, F., Reichart, R., & Korhonen, A. (2015). “Simlex-999: Evaluating semantic models with (genuine) similarity estimation”. *Computational Linguistics*, 41(4), 665-695.

[4] Liu, M., & Yang, J. (2012). “An improvement of TFIDF weighting in text categorization”. *International proceedings of computer science and information technology*, 44-47.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. “Efficient Estimation of Word Representations in Vector Space”. In Proceedings of Workshop at ICLR, 2013. p. 1301-3781.

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). “Distributed representations of words and phrases and their compositionality”. *In Advances in neural information processing systems* (pp. 3111-3119).

[7] Naili, M., Habacha, A. C., & Ghezala, H. H. B. (2016, April). “Parameters driving effectiveness of LSA on topic segmentation”. *In International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 560-572). *Springer, Cham.*

[8] Pennington, J., Socher, R., & Manning, C. (2014). “Glove: Global vectors for word representation”. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[9] Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011, March). “A word at a time: computing word relatedness using temporal semantic analysis”. *In Proceedings of the 20th international conference on World wide web* (pp. 337-346). ACM.

[10] Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). “Aravec: A set of arabic word embedding models for use in arabic nlp”. *Procedia Computer Science*, 117, 256-265.

[11] http://news-eleven.com/

[12] http://thithtoolwin.mmbloggers.com

[13] https://boilerpipe-web.appspot.com

[14] https://en.wikipedia.org/wiki/Tf-idf

[15] https://github.com/ye-kyaw-thu/sylbreak

[16]https://raw.githubusercontent.com/lwinmoe/segment/master/burmese-word-list.txt

[17]https://thanlwinsoft.github.io/www.thanlwinsoft.org/ThanwinSoft/MyanmarUnicode/Conversion/myanmarConverter.html

[18]“Myanmar Sagar Pariyaral Kyam”. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar, February 2017.

Software Engineering

Software Size Measurement for Embedded System using Class Diagram

Thandar Zaw
University of Information Technology, Yangon, Myanmar
thandarzaw@uit.edu.mm

Swe Zin Hlaing,
Myint Myint Lwin,
University of Information Technology, Yangon, Myanmar
swezin@uit.edu.mm,
myintmyintlwin@uit.edu.mm

Koichiro Ochimizu
University of Information Technology, Yangon, Myanmar
ochimizu@jaist.ac.jp

Abstract

Todays, software size measurement is the essential roles for project management. It can measure before developing in software to get the accurate size of software. Due to obtain the functional size of software, there are many measurement methods that have been recognized as international standard. Some measurement methods are designed for business application software and a few methods are designed for real-time application software. Thus, in order to measure the functional size of embedded system correctly, the well-defined class diagrams notation and well-designed FSM procedures should be used in embedded system. So, *COSMIC FSM is one of the well-known methods of FSM which is suitable to estimate the size of embedded software. This paper proposes UML class design notations which can be used to estimate the size of embedded software. This paper also shows the mapping rules which mapped between the class diagram and COSMIC FSM to measure the functional size of software. Finally, the functional size of software is calculated by using COSMIC FSM.*

Keywords- Common Software Measurement International Consortium (COSMIC FSM), UML class diagram.

1. Introduction

Software sizing is used to estimate the size of software application. Several size estimation methods have been proposed. The most popular size estimating methods are Source Line of Code (SLOC) which is based on size related measures and Function Points which is based on function-related measures. Functional size measurement is an important way of measuring software in the early stages of development when the effort and cost estimation is most needed. Several Function Points measurement was recognized as an international standard. The IFPUG [1], MKII [2], COSMIC [5,6 ,9], NESMA [3] and FISMA [4] were also defined and standardized. Among them, one of the functional size measurement methods is COSMIC FSM which was designed to be applied in various functional domains such as business application domain and real-time application domain.

Many researchers proposed the software estimation methods which are applicable for UML sequence diagram notations and have not proposed for UML class diagram. To address this limitation, this paper proposed the UML class diagram notations to estimate the functional size of software. These notations have been translated to COSMIC by using mapping rules with a simple case study of cooker system. This paper is organized as follows: the second section provides related work; the third section presents the proposed system; the fourth section explains the case study and final section is conclusion and future work.

2. Related Work

In [7], Symons, C. described the COSMIC concepts that can be applied in any real-time software requirements to measure the functional size of real-time software to understand clearly for any software engineer with alarm example. In [8], Soubra, H., et al. proposed the design of the FSM procedure based on the documentation of the mapping of the Simulink concepts to COSMIC concepts for the embedded real-time software system. In [11], Luigi Lavazza., et al proposed the UML that can be used to build models according to the COSMIC measurement rules. Asma Sellami et al. [12] proposed the measurement method for sizing of sequence diagram that can be measured both the functional and structural size at different level of granularity. In [10], the author proposed the automated functional and structural measurement of software size from XML structure of sequence diagram to calculate COSMIC CFP. In [13], the author presented a COSMIC based FSM procedure for RTS (Real-Time embedded systems) designed with SCADE and manually applied the FSM procedure to an aerospace system and also compared the results obtained with automatically by a prototype tool. This paper proposed the UML class design model notation. Then, the mapping rules define which can map between these notation and COSMIC.

3. Proposed System

This section described the proposed method based on the COSMIC that uses the specification expressed UML class diagram notation in embedded system to obtain the functional size of software. The proposed system has three measurement processes which are shown in Fig 1. In the measurement strategy phase, the proposed system is analyzed from the popular notation of UML class diagram. After analyzing the class diagram with COSMIC, the mapping rule apply between these diagram and COSMIC during the mapping phase. The functional size of software is calculated by using the COSMIC in the final state of measurement process.

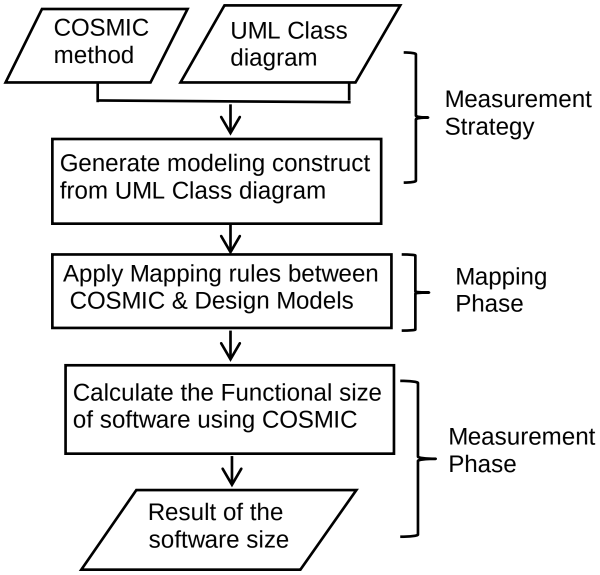

COSMIC method
UML Class diagram
Measurement Strategy
Generate modeling construct from UML Class diagram
Apply Mapping rules between COSMIC & Design Models
Mapping Phase
Calculate the Functional size of software using COSMIC
Measurement Phase
Result of the software size

Figure 1. Proposed System based on UML Class diagram

4. Case Study

The UML use case and class diagram are used in the case study of cooker system. Then the total no of data movements for each class diagram is calculated.

4.1. Functional User Requirement

The paper proposed the specification of a simple version of the cooker system which is used as a case study to determine the counting of COSMIC [13] .Before developing the UML representation, the specification of the cooker system must be defined. The functional user requirements of this system are as follows:

1. The cooker software can get the input from a door sensor and start button. Then it can show the light and heater on/off when the power is switched on.
2. When the start button is pressed and the door is closed, the cooking starts. If the door is open, the start button has no effect.
3. Either the door is open while the cooking is in progress or when cooking is completed, the timer signals will stop.

4.2. Use Case of UML Model

The use case diagram of cooker system is shown in Fig. 2. This system consists of three main functionalities: Pressed Button, Opened Door and End Cooking. The functional users of this system are Door Sensor, Start Button, Timer, Light and Heater.

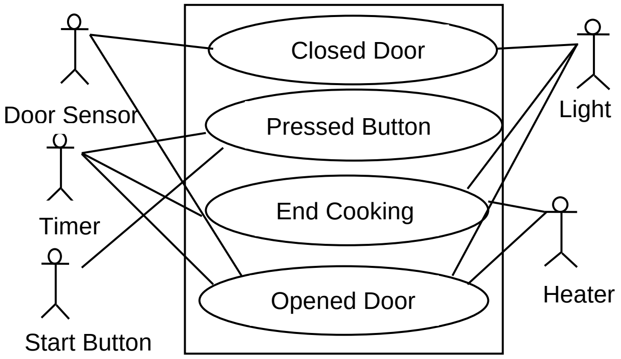


Figure 2. Use case diagram of cooker system

4.3. COSMIC FSM in UML Class diagram

The measurement process of the cooker system mainly comes from the design requirements of UML class diagram. In this system, there are three functional processes of UML class diagrams for the cooker system as shown in Fig. 3 to 5. In Fig. 3, the cooker checks that the door is open or closed. When the door is closed, it sends the signals to start the heater and to switch on the light. The dependency is a relationship between named elements such as class diagram. It is appropriate to identify functional size of class diagram depends on the number and types of messages exchanged. Usage is a dependency in which one named element (client) requires another named element (supplier) for its full definition or implementation. Call is a usage dependency that specifies that the source operation invokes the target operation. This dependency may connect a source operation to any target operation that is within the scope including, but not limited to, operations of the enclosing classifier and operations of other visible classifiers. Call is denoted with the standard stereotype «call» whose source is an operation and whose target is also an operation. Send is a usage dependency whose source is an operation and whose target is a signal, specifying that the source sends the target signal. Send is denoted with the standard stereotype «send».

The types of messages with the corresponding measurement results in CFP units are shown in Table 1. Then, the number of data movements for each class diagram is calculated. Finally, the total size of system is calculated by aggregating all these data movements.

Table 1 Message types in a UML class diagram

Usage dependency	Symbol	Standard stereotype	Data Movement	CFP units
Call	⇢	<<call>	R or W	1
Send	⇢	<<send>>	E or X	1

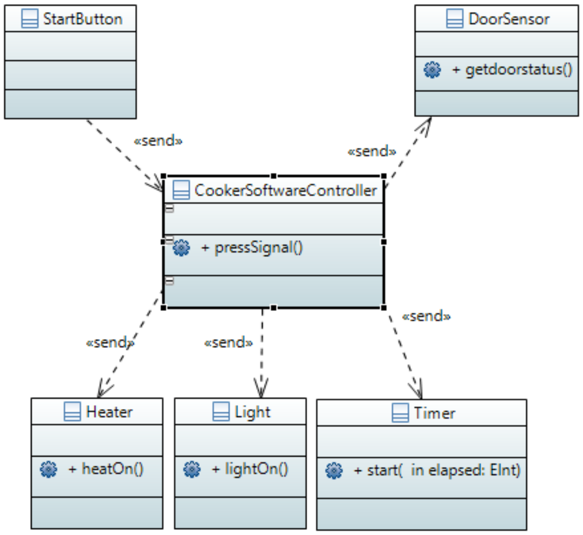


Figure 3. Button Pressed of UML Class diagram

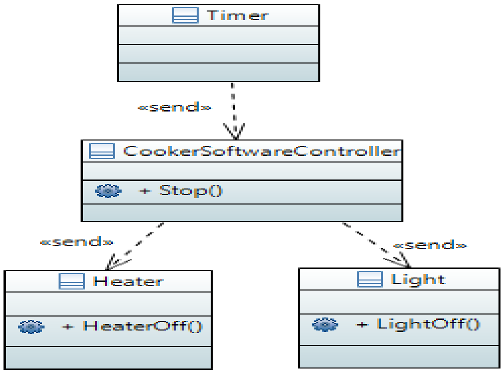


Figure 4. End Cook of UML Class diagram

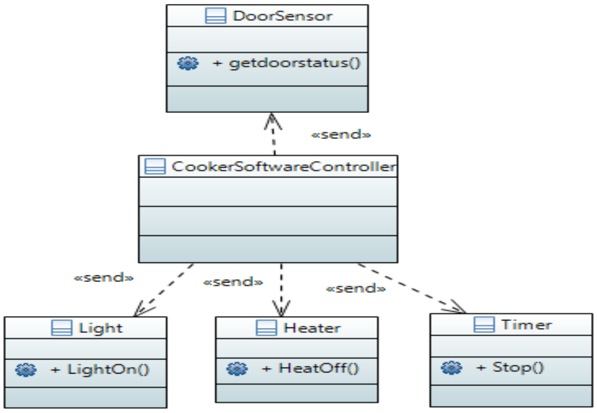


Figure 5. Open Door of UML Class diagram

4.4. Mapping Rules

In this section, the key concepts of COSMIC are mapped to the key concepts of UML class diagram notation.

The mapping rules are described as follows:

Rule 1: Identify the boundary.

The boundary is represented in use case diagram. It shows the application border that is established by identifying the external elements and application system.

Rule 2: Identify the functional user.

The functional user is a class that represents an object or a set of objects that share a common structure and behavior.

Rule 3: Identify functional process.

The functional process identifies use cases in the system.

Rule 4: Identify data groups

The data groups identify that trigger event carries data between objects.

Rule 5: Identify four data movements.

The four data movements are identified as follows:

Rule 5.1: The Entry data movement identify from Rule 2 to Rule 3.

Rule 5.2: The Exit data movement identify from Rule 3 to Rule 2.

Rule 5.3: The Read/Write data movement identify messages that send into or out of the internal persistent storage.
Rule 6: Apply the COSMIC measurement function.
According to COSMIC measurement, each of the data movement in each functional process is added to get the functional size of that process.
Rule 7: Aggregate the functional size measurements.
Aggregate all of the data movements of the functional processes of the whole system into a single functional size value to obtain the functional size of the system.

4.5. Measurement Phase

After defining the mapping rules, the data movements of each functional process have been identified as shown in Fig. 3 to 6. According to COSMIC standard, 1CFP is defined as the size of one data movement.

For the case study, there are three functional processes such as Push Button, End Cook and Open Door. The functional process of Push Button identified 2 Entry data movements. The Entry data movements counted the Push Signal attribute from start button and the Getdoorstatus from DoorSensor. It also identified 3 Exit data movements. The Exit data movements also counted the HeaterOn attribute to Heater, the LightOn attribute to Light and the Start attribute to Timer respectively. The subtotal of functional size for that functional process is 5CFP. The total size of each function is 12CFP by adding all number of data movements. The data movement of each sequence diagram in this system is as shown in Table 2.

5. Conclusion and Future Work

FSM, an important component of a software project, provides information for estimating the effort required to develop the measured software. The early prediction of the size of the embedded software can be achieved in the developmental process as the software development costs are increasing. In this paper, we proposed the UML class diagram notation with COSMIC FSM which is applicable to the case study in cooker system. Then, the mapping rules between these notation and COSMIC FSM define to support the functional size measurement of software. It has been intended to propose the various notations by extending the COSMIC rules.

Table 2. Measurement of data movements for cooker system

Process	Message Sending		Data Movements	CFP
	Message	Component of object involved		
Button Pressed	PressSignal( )	Start button	Entry	5 CFP
	Getdoorstatus (Close)	Door Sensor	Entry	
	HeatOn( )	To Heater	Exit	
	LightOn( )	To Light	Exit	
	Start Cooking( )	To Timer	Exit	
End Cook	Stop( )	Timer	Entry	3 CFP
	HeatOff( )	Heater	Exit	
	Lightoff( )	Light	Exit	
Door Opened	Doorstatus (open)	Doorsensor	Entry	4 CFP
	LightOn( )	Light	Exit	
	HeatOff( )	Heat	Exit	
	Stop( )	Timer	Exit	
			Total	12 CFP

6. References

[1] ISO/IEC: ISO/IEC 20926, Software Engineering- IFPUG 4.1 Unadjusted Functional Size Measurement Method- Counting Practices Manual (2009).

[2] ISO/IEC: ISO/IEC 20968, Software Engineering- Mk II Function Point Analysis- Counting Practices Manual (2002).

[3] ISO/IEC: ISO/IEC 24570, Software Engineering- NESMA Function Size Measurement Method version 2.1 – Definitions and Counting Guidelines for the application of Function Point Analysis (2005).

[4] ISO/IEC: ISO/IEC 29881, Information technology - Software and systems engineering - FiSMA 1.1 functional size measurement method (2008).

[5] COSMIC. The COSMIC Functional Size Measurement Method Version 3.0.1, Measurement Manual (The COSMIC Implementation Guide for ISO/IEC 19761: 2003).

[6] COSMIC. The COSMIC Functional Size Measurement Method Version 4.0: Measurement Manual (The COSMIC Implementation Guide for ISO/IEC 19761: 2011).

[7] Symons, C.: “Sizing and Estimating for Real-time Software – the COSMIC-FFP method”. In: DOD Software Tech News’,

Editor: Data & Analysis Center for Software, USA DOD, Rome NY, vol. 9(3), pp. 5–11 (2006).

[8] Soubra, H., Abran, A. , Stern, S. , Ramdan-Cherif, A., “Design of a Functional Size Measurement Procedure for Real-Time Embedded Software Requirements Expressed using the Simulink Model”, Software Measurement, 2011 Joint Conference of the 21st Int'l Workshop on and 6th Int'l Conference on Software Process and Product Measurement (IWSM-MENSURA).

[9] The ‘‘COSMIC Functional Size Measurement Method, version 4.0: Guideline for Sizing Real-time Real-Time Embedded Software”, 2016.

[10] Meiliana etal. ,”Automating Functional and Structural Software Size Measurement based on XML Structure of UML Sequence Diagram ”, 2017 IEEE International Conference on Cybernetics and Computational Intelligence 20-22 Nov. 2017.

[11] Luigi Lavazza and Vieri Del Bianco, “A Case Study in COSMIC Functional Size Measurement: the Rice Cooker Revisited”, IWSM/Mensura 2009.

[12] A. Sellami. etal, “A measurement method for sizing the structure of UML sequence diagrams”, Information and Software Technology 59, 2015.

[13] Hassan Soubra, Laury Jacot and Steven Lemaire, “Manual and Automated Functional Size Measurement of an Aerospace Real Time Embedded System: A Case Study Based on SCADE and on COSMIC ISO 19761”, International Journal of Engineering Research and Science & Technology, vol.4, No. 2, May 2015.